DOCUMENT RESUME

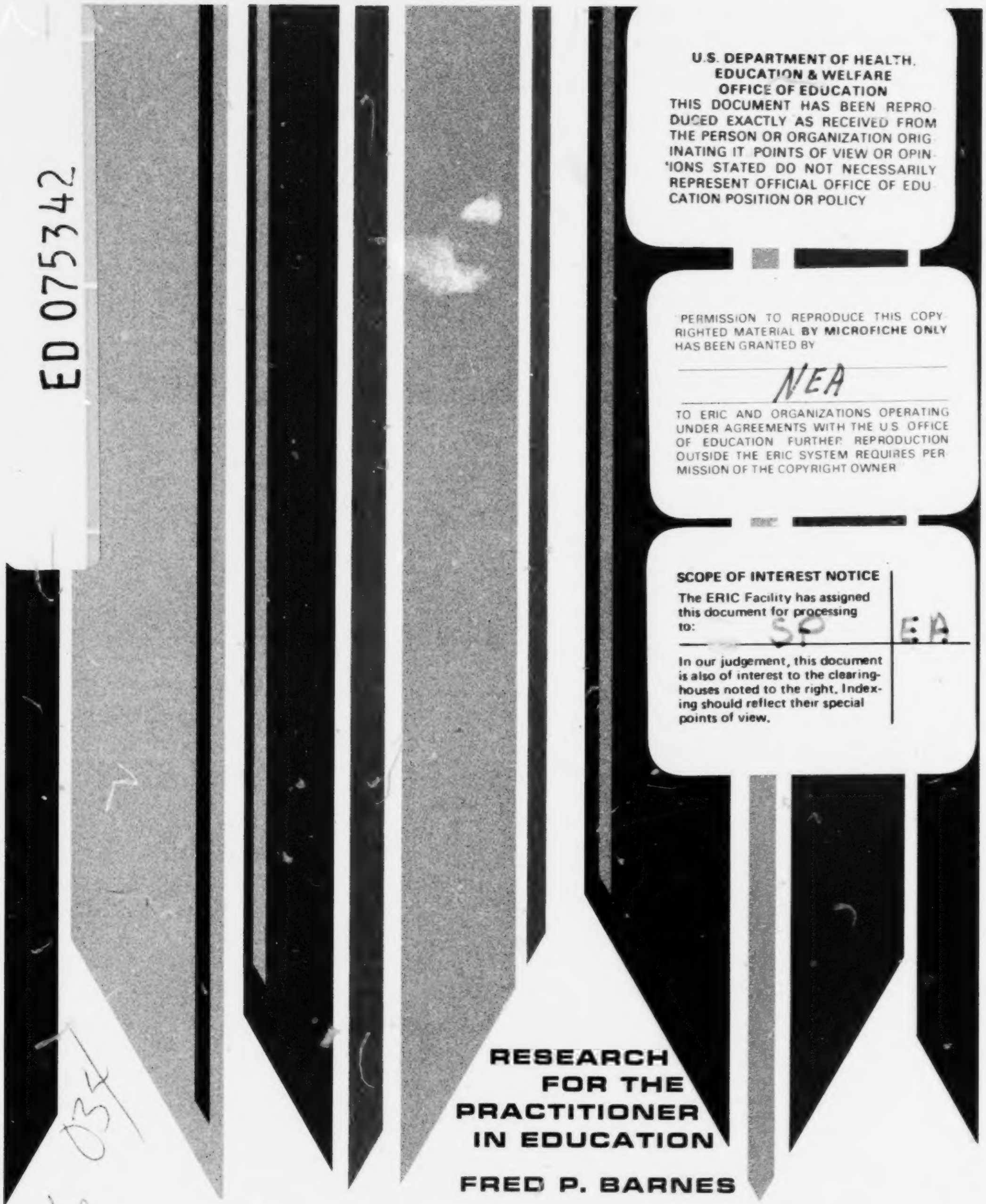
ED 075 342 SP 006 034

AUTHOR	Barnes, Fred P.
TITLE	Research for the Practitioner in Education.
INSTITUTION	National Association of Elementary School Principals, Washington, D.C.
PUB DATE	72
NOTE	148p.
AVAILABLE FROM	National Association for Elementary School Principals, 1801 N. Moore Street, Arlington, Va. 22209 ($5.00 paper, $6.50 cloth single copies)
EDRS PRICE	MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS	Control Groups; Literature Guides; *Research Criteria; *Research Design; *Research Methodology; *Research Problems; *Statistical Analysis; Statistical Data; Test Interpretation

ABSTRACT

This book is designated for an audience of the practitioners of education, especially those located in elementary schools. It is designed to investigate and explain research methods, terms, and concepts so that the reader can narrow down the many ways things are known or said to be known. Included in this book are chapters on the research process in general, problems and hypotheses, variables and probability, statistical tests, analyses, and statistics and the research literature. (JA)

ED 075342



U.S. DEPARTMENT OF HEALTH.
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY
NEA
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER"

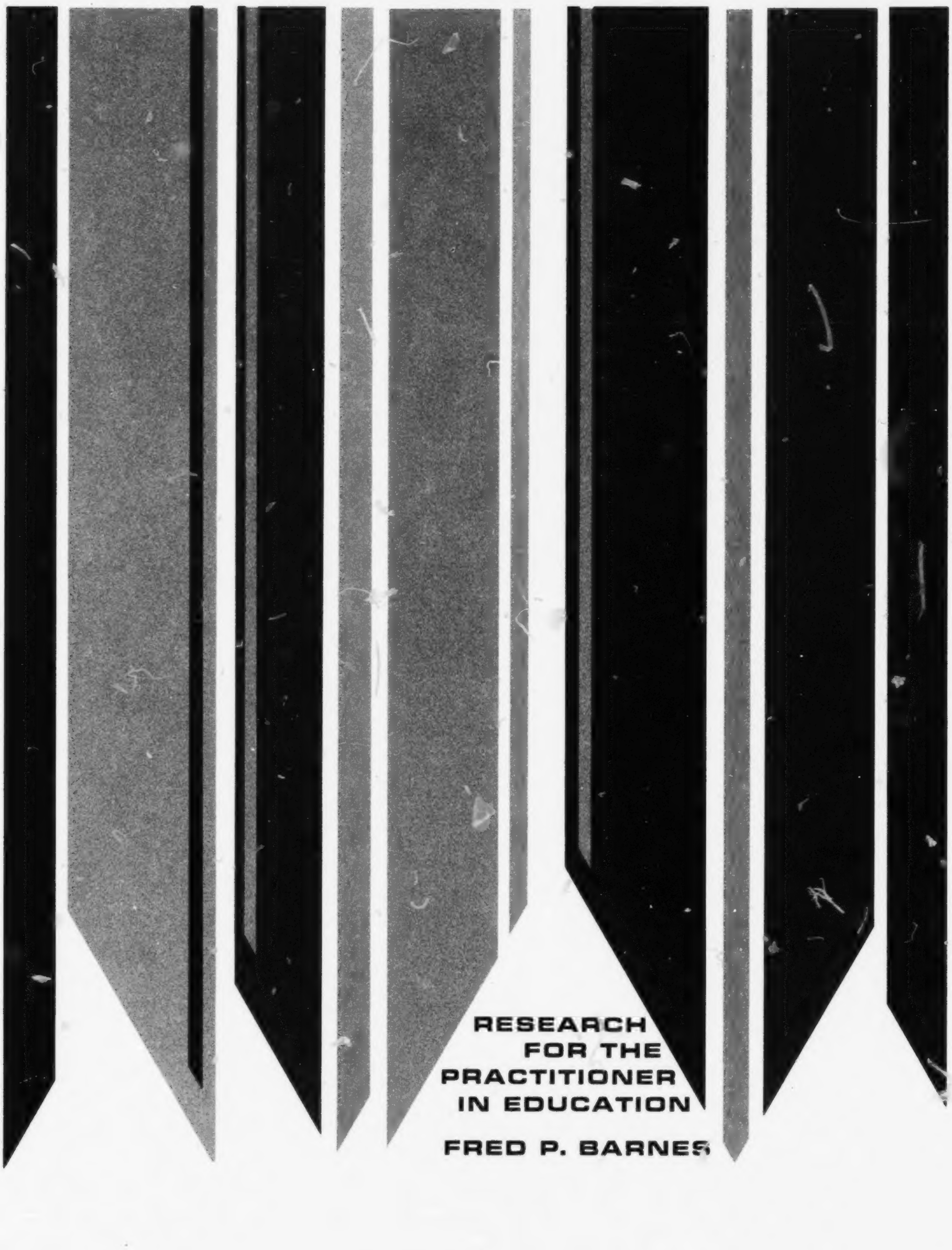
SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: SP | EA
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

RESEARCH FOR THE PRACTITIONER IN EDUCATION

FRED P. BARNES

SP 006 034

FILMED FROM BEST AVAILABLE COPY



RESEARCH FOR THE PRACTITIONER IN EDUCATION

FRED P. BARNES

National Association of Elementary School Principals

Copyright © 1964,
Department of Elementary School Principals,
National Education Association.
All rights reserved.

No part of this book may be reproduced in any form
without written permission from the publisher,
except by a reviewer who wishes to quote brief passages.

Library of Congress catalog card number: 64-22693

Single copies $5.00, paper; $6.50, cloth
2-9 copies, 10% discount; 10 or more copies, 20% discount

Published by the National Association of Elementary School Principals
1801 North Moore Street, Arlington, Virginia 22209

Printed in the United States of America
Reprinted 1967, 1970, 1972

CONTENTS

Foreword vii

CHAPTER 1 **THE SEARCH FOR KNOWLEDGE** 2
Alternative Sources of Knowledge 5

2 **THE RESEARCH PROCESS** 12
Characteristics of Experimental Research 13
Two Ways of Reasoning 18
Seven Steps in the Research Process 22

3 **PROBLEMS AND HYPOTHESES** 26
Researchable Problems 27
Development of Hypotheses 29

4 **POPULATIONS AND SAMPLES** 33
Selecting a Representative Sample 36
Special Problems in Schools 39

5 **VARIABLES AND PROBABILITY** 42
Independent and Dependent Variables 43
Statistical Tests of Probability 46
An Illustration 47

6 **DESIGNS AND DECISIONS** 52
Design 1: "Before-After" Study with Control Groups 53
Design 2: "After-Only" Study with Control Group 58
Design 3: "Before-After" Study with a Single Group 61
Design 4: "After-Only" Study Without Controls 65
Design 5: Ex Post Facto Experiment 68

7 **ANALYSES AND STATISTICS** 74
Nonparametric and Parametric Methods 75
Measurement Scales 78
Stating and Evaluating the Null Hypothesis 79

CONTENTS

CHAPTER 8 **STATISTICAL TESTS** 85
The One Sample Condition 87
The Two Related Samples Condition 89
The Two Independent Samples Condition 92
The More Than Two Related Samples Condition 98
The More Than Two Independent Samples Condition 100
The Measure of Correlation Condition 105

9 **REPORTING RESULTS** 108

10 **THE RESEARCH LITERATURE** 112
Books 114
Encyclopedias 116
Literature Guides 116
Abstracting and Indexing Journals 118
Review Journals 119
Research Journals 119
General Journals Also Containing Research Projects 124
Monographs 127

11 **STAFF TRAINING AND FOLLOW-UP** 129
Some Experiments in Research Training 130
Planning for Local Research Training 131

Footnotes 138

Index 139

FOREWORD

WHEN developing a speaking acquaintance with a book, understanding is sped if one can peer into the book's intent and recognize one's own place in that intent. Sensibly, a potential reader can judge whether the book's companionship will suit his own curiosities and compulsions or whether he will save time and stress by seeking other, more compatible literature.

This book was written for a particular audience: those administrators and teachers on-the-job in American schools—the practitioners of education, located primarily in elementary schools.

So far into the sixties, the schools have been caught up in a ferment of developments and ideas which appear to promise rather sweeping changes for the schools. Only a few of these developments have been systematically tested in school situations, and the professional literature is filled with proposals for experimentation and research studies designed to support or discredit the changes on scientific grounds. By the time that innovations have reached into the schools, it has already become the practitioner's job to evaluate them if he is going to defend his judgments and decisions.

But those who call for research studies in the schools to accompany the changes which are confronting administrators and teachers have given notably scant heed to preparing materials for practitioners which might help ready them for the research tasks which are implied. Quite a gap is left between the need for research in the schools and workable approaches to accomplishing that research in reality.

Perhaps this book can at least partially help to fill that gap. If you are an educational practitioner in a school system, or if you are a practitioner in training in a college program, or if you are a college professor working with practitioners, this book may have direct ideas for you.

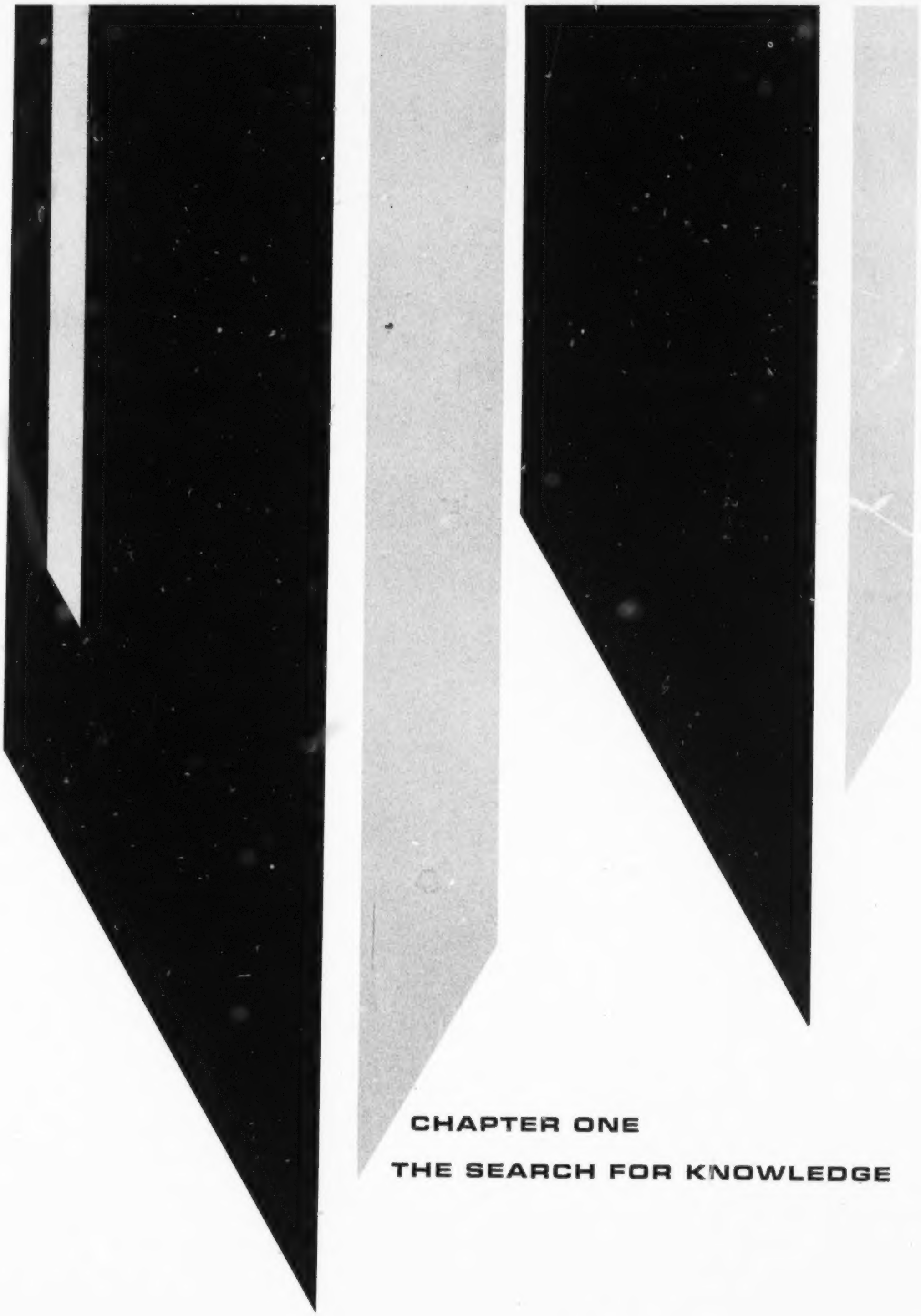
Fred P. Barnes

ABOUT THE AUTHOR

The late Fred P. Barnes was Professor of Education at the University of Illinois where his major fields of teaching were instructional supervision and elementary school administration. His professional experience included eleven years as a classroom teacher in the public schools at both the elementary and secondary levels.

Dr. Barnes is the author of **Practical Research Processes** and **Research Methods in Education** and contributed to **Modern Methods in Elementary Education** and **Improving the Quality of Public School Programs.** He was senior author of **Practical Research Projects,** published by the Illinois Association for Supervision and Curriculum Development, and was a frequent contributor to professional journals. Dr. Barnes received his B.S. and M.A. degrees from Northwestern University and his Ed.D. degree from Columbia University.

RESEARCH FOR
THE PRACTITIONER
IN EDUCATION



CHAPTER ONE

THE SEARCH FOR KNOWLEDGE

Within the teaching profession, and outside of it, too, the talk about education is flavored with terms that would have seemed foreign a scant decade ago. Modern mathematics, teaching machines, electronic language laboratories, team teaching, the nongraded primary—such is the current vocabulary of education, a vocabulary born out of the quest to find better ways of teaching.

In the midst of all this ferment, principals and teachers often find themselves in a difficult spot. How workable are the proposals for innovation? Do the new devices and the new patterns of teaching produce better results than some of our time-honored procedures? Such questions cannot be thrust aside; the practitioner needs answers. He needs answers for himself in making decisions about instruction, and he needs answers for use with parents and other laymen suddenly conversant with the terminology and concepts of educational innovation. But when the practitioner seeks the answers he needs, when he turns to the professional researchers and students for tested evidence, he is often confronted with nothing more convincing than research findings of "no significant difference" between the results of traditional teaching and new ways of teaching.

The practitioner finds "no significant difference" almost everywhere he turns. For example, the *Review of Educational Research* for April 1962 reviewed research on technological media completed over the preceding six years. The over-all verdict about instructional television was "no significant differences" when televised instruction was compared with ordinary classroom instruction. In regard to self-teaching devices and programed materials (teaching machines), the most popular finding again was "no significant differences" compared with ordinary instruction. Language laboratories came out somewhat better, chiefly because the amount of research being done was so sparse, and the review ended with the observation that much more controlled research is vitally needed.

It would be a mistake to interpret the "no significant difference" findings as either a validation or a condemnation of certain technologies applied to classroom instruction. Rather, the findings may indicate that the media have been used unimaginatively or that the research has been guided by the wrong questions. Nevertheless, repetitive research results of this sort tend to build up the suspicion that neither television nor programing, for example, has plowed up evidence to indicate that it is a vehicle on which teachers might ride to newer and higher levels in teaching—at least not yet.

Another illustration of the lack of convincing research results is found in the area of school organization. The self-contained elementary school classroom, heterogeneously grouped, has recently come in for re-examination, in part because of renewed interest in departmentalization and ability (X, Y, Z) grouping. Again, principals and teachers are faced with the "no significant difference" puzzle. Departmentalized schemes like the Dual Progress Plan and the Amidon School are still leagues away from an impartial demonstration that they are achieving a significant difference in children's learning as compared with ordinary organizational patterns. In relation to grouping, the NEA's publication, *The Principals Look at the Schools,*[1] revealed a marked trend, just since 1957, away from heterogeneous and toward ability grouping in American schools. This trend must reflect a halo movement since research studies over the past twenty-five years have produced only indeterminant findings which have clearly fallen short of demonstrating any alleged superiority of grouping according to ability.

Some of the standard references on elementary school education present similar research findings and then hasten to the generalization that administrative schemes—for example, the grouping of pupils and the organization of subject matter—cannot be expected to solve truly important instructional problems anyway. Of course, this is simply a neat piece of rationalization. It accomplishes nothing more than to dismiss the principal from responsibility and push the onus for instructional improvement back onto the teachers where it can serve its customary divisive function. More important, this side-stepping of the question also side-steps cogent ideas about what *can* be expected to solve important instructional problems.

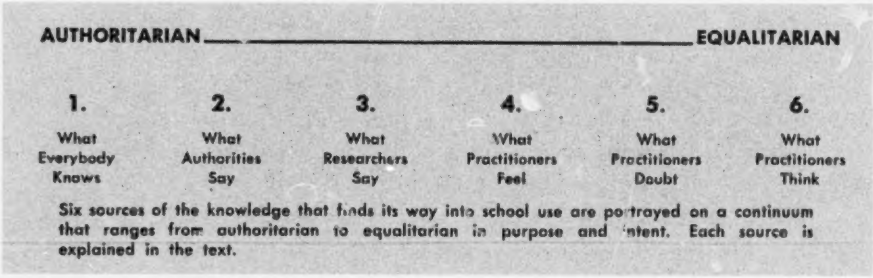
Principals and teachers cannot avoid decisions on instructional problems. Neither can they rationalize gaps in their knowledge by becoming members of the not-so-exclusive club of "no significant differences." They are, after all, engaged in a profession which deals with intellection as its stock in trade. Small wonder, then, that as purveyors of knowledge, they are expected to apply the fruits of knowledge to the conduct of their specialty. This is the expectation generally held by the public, and it is both reasonable and complimentary. For that matter, educators themselves abjure lack of knowledge, particularly lack of knowledge about their own bailiwick.

Quite understandably, then, principals and teachers look for ways to explain and make sensible their own instructional practices; any ostensible lack of knowledge is highly unconvincing to everyone concerned. The need for knowledge becomes particularly acute when a school is involved in one of the "new" practices such as televised teaching, programed instruction, departmentalization, or ability grouping. In such instances, both the public and the practitioner are especially anxious for evidence of effectiveness, yet the relevant knowledge is desultory to say the least. But decisions must perforce be made, and it is often hard to tell whether some specific practice got into a school via intelligence or salesmanship.

Whether their practices have the glitter of the new or the comfortable assurance of the customary, it is a characteristic of American schools to be constantly searching for ways to improve. If the requisite knowledge can't be found in research evidence filled with "no significant differences," practitioners look elsewhere for answers which will seem satisfactory to themselves and to their public.

ALTERNATIVE SOURCES OF KNOWLEDGE

What are the alternative sources to which principals and teachers may turn for the information they need in making instructional decisions? Based on a sort of continuum, there seem to be six approaches to the quest for new knowledge or for validation of information already known. These approaches range along a continuum from the authoritarian to the equalitarian, from dependence upon the traditional knowledge of "everybody" to relative independence and self-reliance in creating and testing ideas.



Six sources of the knowledge that finds its way into school use are portrayed on a continuum that ranges from authoritarian to equalitarian in purpose and intent. Each source is explained in the text.

Figure 1. Sources of Knowledge for School People

What everybody knows. The most convenient and the most widely used place to look for defenses of what we do in the schools is in "what everybody knows." We all know what "everybody" knows, so it takes little effort to find rationales for what we are doing. We can begin and end our search for knowledge with the assumptions of common parlance. For example: subject matter consists of what is found in textbooks; younger students should not and cannot learn the same things as older students; quick starters and hard workers will learn more and faster in school than late bloomers and rebels; mental discipline is one of the objectives of teaching. Here is the authority of the majority.

This approach to the search for knowledge, with its efficiency and directness, would be economical and preferable if we could only be confident that what everybody knows is true. But such is not the case. It may not have been true to begin with and, in any event, knowledge does not stand still. Recent advances in the various fields of knowledge have made obsolete much of the subject matter contained in the textbooks; psychological research studies have hinted that very small children can learn and are learning very big things; late bloomers and rebels may in time surpass fast starters and hard workers; and mental discipline was discredited on many fronts several years ago.

Taking the easy way in the search for knowledge and depending upon our favorite assumptions eventually becomes a serious hindrance in a world of expanding ideas. It hitches us to the past. And once hitched, there is no way out, because the seeds of fresh thought are not to be found in what everybody already knows.

What authorities say. The next source of knowledge is also one of the favorite founts of knowledge about school practices. The literature of education is filled with the conclusions, blandishments, and analyses of educational "leaders." The intent of this literature is simple and direct enough: if principals and teachers will read or hear the ideas of outstanding scholars in the field, they will be prepared to transfer those ideas to the practical problems of teaching children. What an authority says is accepted as valid simply because an authority said it.

This avenue to workable knowledge, while sanctioned by the academic world, is filled with unsuspected pitfalls for both the practitioner and the authority. One problem is communication. To use the language of information theory, the message transmitted by the sender (the authority) may be woefully garbled by the receiver (the practitioner). Certainly this was the case with John Dewey's books on education. The receivers of his messages came away with many more and fanciful interpretations of his ideas than he ever intended. In the same way, intrusion of static may be deflecting Conant's messages on the American high school. In Conant's name, a few hard-to-believe changes are creeping into some of our secondary schools.

To read or hear an authoritative dictum and then attempt to put it directly into use without empirical testing is risky business. It is risky in the first place because it involves trafficking in other people's thoughts and prestige. It is risky in the second place because it is easy to misinterpret the dictum and be misled into doing the wrong things. And most important, it is risky because the chain of thought—from the authoritative production of ideas to the consumption of those ideas—includes no interim sequence of challenging and testing.

All of this is not to imply that other people's ideas have only questionable use in the search for knowledge. They have great value if they are used as motivators for one's own ideas and if they are thought of as propositions which must be tested before they are accepted as true. But our very reverence for the authority figure in education ordinarily prevents our doubting him even in the absence of first-hand evidence, and it is always faster and easier to use an authority's name in lieu of evidence.

What researchers say. Increasingly, a show of being acquainted with "what research says" is coming to be a popular mark of unassailable sophistication in the educational world. Certainly, an affinity for knowledge derived from research bespeaks a preference for objectivity and fact. The reports of researchers represent a kind of sanitized information uncontaminated by human preference or whim; decisions are reached through processes quite independent of the frailties of human thought.

When research is sanctified in this way, it often exerts an influence which is not very different from the effect of pronouncements by authorities. The chief difference resides in the fact that the researchers have been substituted for the authorities. The practitioner—the *consumer* of research—is simply getting the word from the researcher—the *producer* of research—rather than from the authority. But this difference is not sufficient cause to suppose that one source is any less authoritarian than the other.

The researcher himself, when occupied with his work, probably offers the best evidence that this purified view of research is nonscientific. The competent researcher would never approach the reports of other researchers as superhuman works quite insulated from humanity. He would be more likely to regard them as tentative glimpses into another human being's way of thinking, glimpses which should be carefully criticized and challenged. Most important, he would probably recognize that the chief value of someone else's researches lies in the clues they may offer for further studies of his own. He would certainly not think of them as the final word on any subject in which he was interested.

As yet, we have a considerable distance to go before we develop the wit to use research in the same scientific way that it is produced. It seems that we have fallen into a notion that research is for export, that it offers conclusive answers, when we should be using it to provide clues to our own search for understanding. If we allow the findings of research to restrict, rather than liberate, our ways of thinking about educational problems, then school people will lose the potential of personal reward that comes from being involved in scientific investigation, and there will be no benefit to the vitality of their professional knowledge.

What practitioners feel. The sensate impressions and consequent actions of school people seldom find their way into the literature of education, yet in some cases they represent a valued source of knowledge. There are some teachers who possess an unerring instinct for workable, effective educational practices. And there are some administrators who naturally seem to sniff out ingenious procedures for school operation. These are the intuitive practitioners who intellectually leap their way into stunningly impressive accomplishments without quite being able to rationalize or verbalize why they do what they do. These are the people who, in the language of aeronautics, "fly by the seat of their pants." While their colleagues are pondering, studying, and debating, these naturals seem to arrive easily at insights and achievements that the others can't quite manage.

The intuitive practitioner can be found in all professions. A particular physician or attorney may exhibit a kind of class in his practice that seems to escape most of his colleagues, even though their preparation may have been similar. It may be that whatever his profession, the practitioner is uniquely situated to utilize and perfect intuitive knowledge because his practice offers him an immediate field of test and feedback for his ideas; this is certainly not the situation of the cloistered scholar.

In education, the practitioner with built-in radar has long been an intriguing but baffling subject for researchers. If only his behavior and shrewd thought patterns could be described and communicated! But this is an elusive goal; neither the researcher in teacher and administrator training nor the intuitive practitioner himself has yet demonstrated much ability to divulge the secrets. For example, one can consult the *Encyclopedia of Educational Research* or the notably artful teacher on the simple question, "What do talented teachers *do* to produce such a walloping effect on their students?" and come away with the question still intact and the answer still fragmentary.

As a source of knowledge, the practitioner with an instinct for the visceral is charming, but he is not far removed from the authoritarian.

He is somewhat mystical, even to himself, and thus cannot be publicly shared. Wisdom from this source, like the pronouncements from the authority and the researcher, generally is taken on faith rather than on understanding.

What practitioners doubt. In the continuum of sources of knowledge, the first vigorous break with tradition, authority, and mysticism enters with the man from Missouri. The thoughtful skeptic who demands to be shown before he will embrace an idea has taken the first large stride toward mature, independent thinking.

Of course, the persistent freethinker is a troublesome guy to get along with, especially for the authoritarian kind of person. Fortunately, there is a nice, solid cadre of principals and teachers who serve the very useful purpose of irritating the proponents of various educational doctrines and criticisms. We have built up a long-time qualmishness about such people, however, and tend to dismiss them as empiricists, or simply hardheads. Because they tend to remain unmoved, even in the face of commonly accepted educational sanctions and rituals, they are regarded as hard to educate (ask almost any college professor). Yet the practitioner who congenitally questions and presses for the answers is busy stirring up the serene fields of accepted information and therefore is of independent worth in the advancement of knowledge. Philosopher Josiah Royce put the matter this way: "Despise not doubting; it is often the best service thinking men can render their age. Condemn it not; it is often the truest piety."[2]

The conscientious skeptic is curious about other people's ideas. He is almost compulsive about this questioning. He is motivated by the intellectual fun of taking ideas apart to examine them. In the case of the skeptical educational practitioner, this involves trying out a new instructional method or material—such as ability grouping or a programed textbook—before enthusiasm is allowed to set in. The trial can be accomplished through either formal or informal experimentation within the confines of one's own job. After all, any procedure, be it new or old, is recommended to school people because someone believes he has evidence of its preferability for use in certain situations. What, then, is more directly logical than to give the procedure a trial in one of the situations for which it was intended?

The redoubtable, questioning practitioner makes a solid contribution in the search for knowledge, just by pointing to the tough sort of questions that ought to be asked. But of course, something else is really needed. If the doubter only focuses his attention on dissecting other people's ideas, his own original thinking tends to be crowded out of his purview, and he remains stimulating but incomplete. He has taken the first very necessary step toward scientific thinking and investigation, but to be satisfied with just this much is to drink the apéritif and miss the meal.

What practitioners think. The reflective school person, the practitioner who ruminates a good deal on the embattled state of the schools, is sooner or later bound to risk trying his own ideas and making his own mistakes. He will become impatient with being led by second-hand knowledge to the mistakes that are seemingly endemic to schools in flux. This preference for independence eventually blossoms into a full-fledged approach to education based upon original experimentation and research.

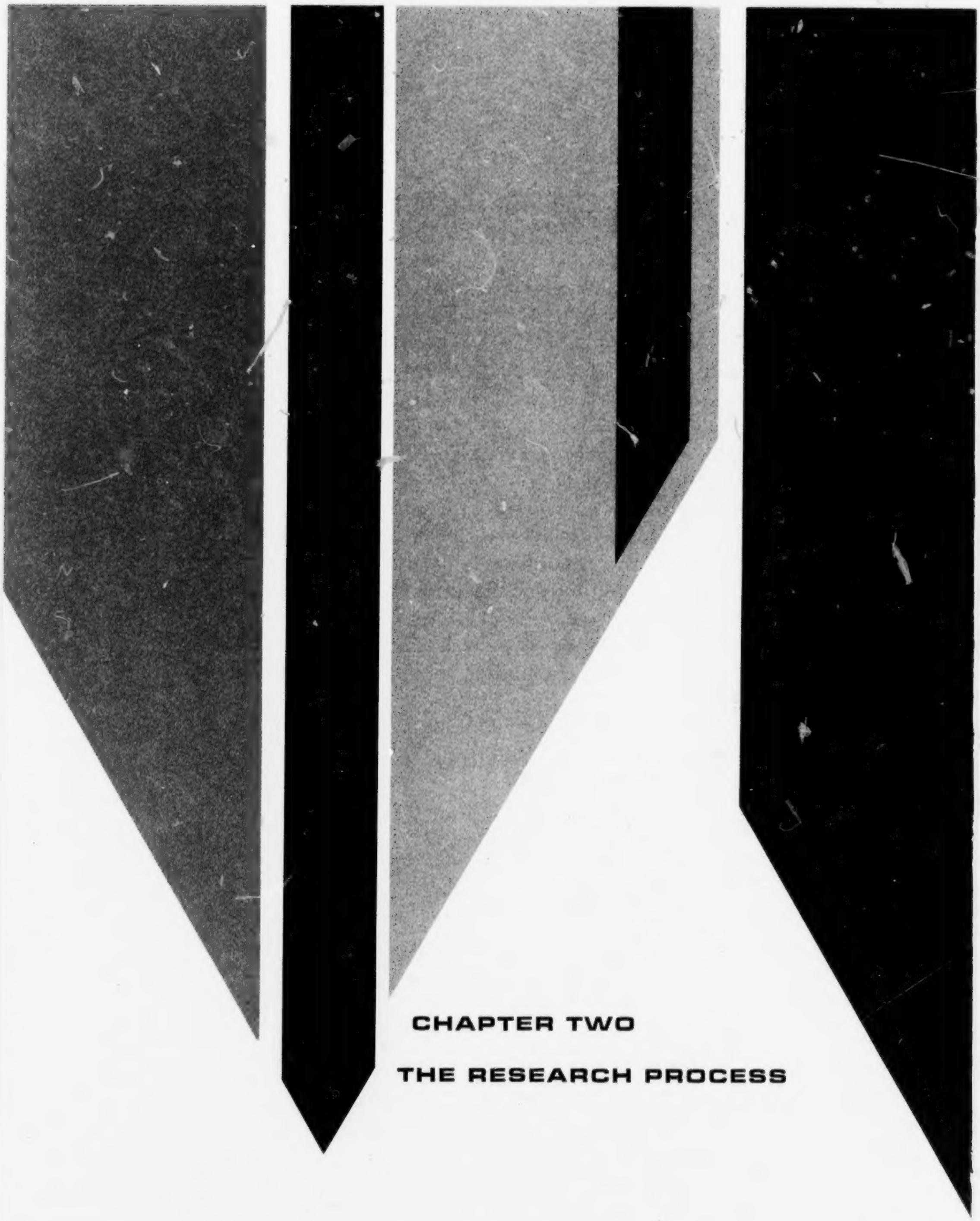
The sense of following one's own reasoning, and then verifying it through rigorous test, is self-fulfilling and personally rewarding. Furthermore, scientific inquiry represents the sole means through which man can enlarge his verified knowledge. It is, therefore, of supreme importance to human enterprises. Proficiency in scientific thinking and research procedures elevates the school practitioner to equality with the intellectual world around him. Authoritative pronouncements become far less commanding, and ideas from all other sources of knowledge fall into place as tentative leads which are perhaps worthy of test.

It would be a serious mistake to allow all of this to sound overpowering and grandiose. Research can be as large or as small, as ambitious or as simple, as anyone wants it to be. And if it is steeped in humility, so much the better. Far from being an inordinately complicated set of distant rituals dressed in a white gown, research is best when it becomes as personal and comfortable as a well-worn suit of clothes with baggy trousers. A very clear statement on this theme was made by scientist Thomas Henry Huxley, and it applies rather neatly to the point being developed here:

"You have all heard it repeated, I dare say, that men of science work by means of induction and deduction, and that by the help of these operations, they, in a sort of sense, wring from Nature certain other things, which are called natural laws, and causes, and that out of these, by some cunning skill of their own, they build up hypotheses and theories. And it is imagined by many, that the operations of the common mind can be by no means compared with these processes, and that they have to be acquired by a sort of special apprenticeship to the craft. To hear all these large words, you would think that the mind of a man of science must be constituted differently from that of his fellow men; but if you will not be frightened by terms, you will discover that you are quite wrong, and that all these terrible apparatus are being used by yourselves every day and every hour of your lives."[3]

Gaining acquaintance and familiarity with Huxley's "terrible apparatus" is an advisable antidote for the mystery and bunkum which surround research activities and their products. Independent thinking and ingenious testing of ideas—together with appropriate methods and techniques for so doing—are not difficult and forbidding. If they were, it is doubtful that a workable scientific content could ever be achieved in education because it is the practitioners who are in direct contact with pupils and hence through them that scientific findings finally reach the classroom. Indeed, principals and teachers are already using the thoughtful methods and feelings of science and research "every day and every hour of [their] lives," often without knowing that they do so.

It is, then, to the reflective school person, to the principals and teachers who wish to find out for themselves, that this book is addressed. In the search for knowledge, the practitioner who understands and utilizes the methods of research will find himself ahead of the game. Rather than being helplessly frustrated by reports of "no significant differences" or mutely obeisant to the pronouncements of authorities, he will have at his command the tools with which to search independently for the knowledge he needs.



CHAPTER TWO

THE RESEARCH PROCESS

Research is a way of dealing with ideas. It is nothing more than this, and it is nothing less. Most research deals with ideas from the standpoint of one or the other of two types of orientation toward time. The first type of research looks from the present to the past; it represents an attempt to explain what has already happened. Historical research and surveys are two examples of research that looks behind us at where we have been. The methods used are chiefly descriptive in nature. The second type of research reverses the orientation toward time and looks from the present to the future; it represents an attempt to test what would happen *if* . . . . Almost all experimental research in the physical, biological, and behavioral sciences is an example of research that looks ahead of us, at where we might be *if*. . . . The methods used are chiefly inferential in nature.

In both historical and experimental research, the ideas of the researcher and the ways in which the ideas are handled constitute the research. The historian attempts as accurately as possible to describe what he thinks has happened; the experimental scientist attempts as precisely as possible to test what he thinks would happen if relevant controlling factors were deliberately altered.

This book is chiefly concerned with experimental research. The experimental research process utilizes certain approaches to thought and investigation which have been developed in order to increase the likelihood that the results of the research will be relevant, sensible, and useful. The essential nature of the research process is discussed in this chapter.

CHARACTERISTICS OF EXPERIMENTAL RESEARCH

Experimental research has certain fundamental characteristics. The ways in which experimental research does and does not attempt to test ideas are evident in the following discussion of ten of these characteristics.

1. Research ideas are restricted by the requirement that they be testable. Almost everyone has an occasional stirring and unusual idea which soon withers for lack of a promising way to try it out. The effort of following an idea or inspiration to its logical conclusion is worthwhile only when the researcher has a reasonable assurance that observation or experimentation in the natural world can provide the needed information.

2. Theories and speculations are closely related to reality. One of the most rigorous tests of a theory is to ask how adequately it explains phenomena in the natural world. Most of us have become accustomed to that fatal dualism, "Maybe it will work in theory, but it won't work in practice." The researcher is likely to insist that a theory which cannot be verified through empirical research and which won't work in practice was probably an incomplete theory to begin with. He is likely to regard a theory as one of the most useful ways to explain practical events. If a theory really is divorced from the actual world, it is not serving its highest function and is most likely only masquerading as theory.

3. Simplicity in ideas and conceptualizations is the ideal. Direct, clear ideas are very difficult to come by and maintain. It almost seems that the more education some people acquire, the more complicated and cluttered their mental machinery becomes. This state of affairs is scarcely an asset when it comes to solving problems. The elegant idea in research is frequently so clear and obvious that others are heard to ask plaintively, "Why didn't I think of that?" And it is the simple, clear idea which very often is the richest and deepest. Who has not been impressed with the elegant formula of Einstein's theory of relativity, $E = MC^2$? Embodied in this seemingly simple formula are the new ideas of space, mass, time, motion, and gravitation that laid the foundation for atomic fission.

4. Research sets out to test, not to prove. "Proof" is a logical impossibility and quite beyond the realm of research. The researcher seeks answers to questions through the application of scientific procedures. He does not think of himself as a supermortal out to discover and prove ultimate truth. Instead, he is likely to regard himself rather humbly as a tester of ideas which may provide clues to the solution of problems. The phrase "research proves" possibly applies to nothing more glorified than toothpaste and aspirin advertisements on television.

5. The concept of "failure" is an archaic interference in research activities. In a very real and useful sense, it is not possible to fail in research. The expected norm for research in all of the sciences is "non-significance." This means that research projects which result in "significant" findings are by far in the minority. In many cases, there is more to be learned from the reasons why some hypothesis failed to reach "significance" than might have been learned from a supposedly successful experiment. The intrusion of the failure concept into research is antiscientific; it encumbers the researcher with implied guilt feelings and with a moral judgment which is quite foreign to research objectives.

6. The potential value of a research project is directly related to the cogency of the questions asked. Research always starts from a question or problem of some sort. There are two general kinds of reasons for asking research questions: intellectual reasons, based on the desire to know simply for the satisfaction of knowing; and practical reasons, based on the desire to know for the sake of being able to do something better or more efficiently. The first kind of question is sometimes seen as leading to "pure" or "basic" research; the second, to "applied" research.

At times, these two types of research are discussed as if they were mutually exclusive and as if one were better than the other. But in actual practice, research on practical problems may lead to discovery of basic principles, and intellectually motivated research may yield knowledge that has immediate practical usefulness. Whether the motivating purpose of a given investigation is primarily intellectual or primarily practical, the requirements of good research procedure are essentially the same.

7. The methods of research are intentionally devised to prevent the researcher's deluding himself and others. It is so easy to believe that facts have been demonstrated, when in reality nothing has been demonstrated at all, that the precautions common to research seem downright negative. For example, in a statistical study, we work with a "null" (or "no difference") hypothesis. If we reject the null hypothesis at the five per cent level, this means that we are willing to be wrong in that decision five times out of a hundred replications because of sheer chance or sampling error. In a more or less complicated research study, the chances are so great that our presumed findings will be confounded and that we might be mistaken in our decisions, that we lean over backward to account for possible errors. We try, so far as possible, to make certain that our questions and procedures are relevant, reliable, and unbiased.

8. Values play a legitimate and important part in research activities. Not too long ago, research was thought to be completely objective, dealing only with facts. Because values are subjective and concerned with human preference, it was taken for granted that they could not be usefully investigated through the methods of research. This may have been true at one time, but the behavioral sciences have now worked out techniques for identifying and comparing values through attitude and opinion scales and the like. Ingenious ways have been developed to transform what seem to be questions of value into questions of fact.

Opposition to continuous promotion through the grades, for example, may be based on values having to do with "upgrading the standards of academic work" or "using the threat of grade failure" as a motivating device. Whether grade failure does or does not result in improved standards, or whether it does or does not provide desirable motivation, is a question of fact and therefore open to answer by research. In other words, the empirical *effects* of values can be systematically studied and inferences drawn concerning the values themselves.

Values are also an important part of research in a quite different context. The researcher has a moral obligation to the subjects of his research. In educational research, the teacher and administrator have an ethical responsibility to students, parents, and the school community. It is no more defensible to employ experimental deceit in research undertakings involving students than it is to hoodwink students in the educational process. Frequent use of what is regarded as deceit (for example, projective type tests which masquerade as ordinary achievement tests) may undermine the necessary confidence between investigator and subject or between teacher and student. Students, as well as researchers, have personal values which are apt to become involved in the research process. To assume that these are freely exploitable is as unreasonable as for a surgeon to approach a healthy man with the request, "Pardon me, sir, may I make a deep incision in the interests of science?"

The prime ethical principle is based upon the recognition that achievement of the investigator's objectives is dependent upon his respect for the subject's values. While the principle is applicable to the entire field of the behavioral sciences, it is particularly crucial when the subjects are a captive audience of school children. Failure to respect the children's values is totally inconsistent with the purposes of education in our society.

9. The methods of analysis, of logical deduction and statistical inference, should fit the limitations inherent in the problem being investigated. It has become quite fashionable in the world of teaching today to borrow—or at least strike the stance of borrowing—ideas from the "basic disciplines." In fact, the "interdisciplinary approach" to school concerns has almost become a panacea for educational ills, despite the fact that the exact meaning of the phrase seems to be considerably more than hazy to those interdisciplinarians who use it. Of course, borrowing from the parent fields (who frequently disavow paternity) has long been a trademark of the schools and of education. But the cost for so doing has been heavy. Too frequently, the schools and students of education have borrowed both the methods and the findings of research in the related fields, only to discover that neither the methods nor the findings fit particular educational problems.

Ralph Tyler, Director of the Center for Advanced Study in the Behavioral Sciences, pointed to the lack of "fit" at the First Annual Phi Delta Kappa Symposium on Educational Research in 1960. "Because of [the] difference in the kinds of questions under study, the behavioral sciences not only may fail to provide direct answers to the questions of primary concern to educational research but they may also contribute less in the nature of research findings than they do to the development of conceptualizations useful in planning and interpreting studies of education and to the invention of techniques for designing studies, and for collecting and analyzing the data of educational research."[4]

This is not to imply that the behavioral sciences lack value for education. Rather, it is to point toward the need for discrimination between the techniques and methods which fit our limitations and those which uncritically do not. For example, education is a purposeful enterprise conducted to accomplish certain social and personal ends. Education is directed through appropriate objectives which make it goal oriented. The basic questions in education tend to be directed toward the determination of objectives; the selection and organization of educational experiences; the selection and guidance of students; the kinds of educational personnel required; administrative organizations, policies, and procedures; selection and use of instructional materials and facilities; and civic support of educational institutions and programs. Logical and philosophical analyses can be instructive when they meet these particular kinds of limitations. When they do not, they tend to vaporize into abstractions.

Similarly, it is highly possible for statistical techniques to squeeze educational problems into mathematical models which distort the real problems of schools. For example, how shall we deal with a small number of cases, such as the twenty-five or fewer pupils in the ordinary classroom, when traditional statistical tests of hypotheses improve as the sample size increases? How can we test, statistically, an obviously non-normal group of children when statistical models tend to be based on the normal distribution? How do we interpret research that involves huge groups of children?

These problems, and ones like them, tend to sound quite defeating. Fortunately, research designs and analyses which come very close to fitting the limitations currently present in actual school problems have been worked out, and new ones are being invented every day.

10. The researcher courts recognition through the power of his tested ideas, not through the attractiveness of his rhetoric. Eventually, most investigators find it necessary to shift the focus of their attention from analysis of data to written communication of their findings. When this occurs, the researcher becomes a writer, but he is still bound by the same careful rules intended to avoid self-deception that characterize the conduct of research itself.

Unlike other writers, the scientific author is not free to choose what he will include and what he will leave out on the basis of the effect he wishes to create. The basic rule which the investigator must observe is to give all the evidence that is relevant to the research question asked. This he must do, whether or not the results are in accord with his own views.

To do a good job, the research writer will tell enough about the study to enable readers to judge the adequacy of its methods, to form an opinion of how seriously the findings are to be taken, and to repeat the study with other subjects if they wish. Crucial points frequently included are the study design; the method of identifying and treating the variables; the nature of the sample; the data collection techniques; the method of statistical analysis; and the specific levels of confidence accepted in deciding whether differences are to be considered significant. This kind of detailed writing lays its claim to distinction in that it invites the reader to make his own independent check of the procedures used and conclusions reached.

Preferably, the research writer will be very careful not to claim more than his data and analyses will allow. Other writers may dramatically clinch an argument by the trick in logic of generalizing from a particular case to an entire universe: "Johnny can't read; therefore, most children can't read." But the researcher only discredits himself and his work if he indulges in overgeneralizations and pre-ordered conclusions.

TWO WAYS OF REASONING

Students of semantics have fairly well established the fact that our ideas, concepts, and thoughts depend greatly upon the language we use. A person who speaks a language that has an entirely different structure from that of English—such as Japanese, Chinese, or Turkish—does not even think the same thoughts as an English-speaking person. Because the research process deals primarily with the creation and manipulation of ideas, the researcher needs some consciousness of semantic determinants and their implications for reasoning.

Logical-verbal reasoning. The thoughts we have depend largely upon the words we use. Our words are verbalizations of the ideas they represent; as such, they serve as clues to the meanings we seek. Dependence upon verbal clues means dependence on a highly developed, subtle, and complicated form of symbolism. Through the centuries, human beings have agreed to let certain sounds (speech) and certain marks (writing or printing) stand for specified happenings and phenomena. The *symbolic process* is the process through which human beings arbitrarily make certain things stand for other things.

As human beings, we can manufacture, manipulate, and assign values to our symbols. For example, we can agree to let X stand for the boys in a school and Y for the girls. Then we can let the symbol N stand for all the X's and Y's. N becomes a symbol of symbols. It is important to note that there is no necessary connection between the symbol and what it symbolizes. This is true of all our words and phrases. The word *horse* bears no resemblance to the characteristics of the animal on the racetrack, the cowpony on the range, or the Shetland on the farm. It is a symbol—an abstraction—which expresses what is common to the racehorse, the cowpony, and the Shetland, ignoring the differences among them.

The symbolic process is made possible through the *process of abstracting*. This process is an indispensable tool with which we select certain characteristics of an object or event that suit our purposes and leave out other characteristics. Our words, or verbal symbols, are a form of shorthand. They are abstractions which permit us to hold problems in place while we work on them. They permit us to compare situations by examining likenesses or differences. They also make it possible to build conceptions which have no direct counterpart in the world around us. In fact, everything that we know is an abstraction.

Some of our abstractions are at a higher, more inclusive level than others. Highly abstract terms and phrases ultimately serve as classifications and generalizations. Terms or phrases at lower levels of abstraction help to make these generalizations clearer and more specific. For example, the generalized concepts of "work," "play," "communicate," and "read" are at a relatively high level of abstraction. They stand for characteristics that a multitude of activities have in common and leave out most of the specific characteristics of these activities. The abstraction "work" does not necessarily imply rowing a boat, building a garage, or managing a business, yet all of these activities may be work. "Play" might mean anything from having a game of checkers to indulging in pun-making. "Communicate" could vary from one person's sly wink at another to sending a telegram, and "read" might mean anything from interpreting a musical score to receiving Morse code over a telegraph.

The test of an abstraction is not whether it is high or low level, but whether higher abstractions can be meaningfully related to lower ones in the same system of relationships: whether highly abstract ideas can be empirically tested at lower, more tangible levels. Nor is theoretical thinking of a higher type than practical thinking. A person who has command of both theoretical and practical thinking and can easily alternate from one to the other is better able to plan, to invent, to predict, and to test than he who commands only one type of thought.

Statistical reasoning. To this point we have been considering a logical-verbal approach to the research process, embedded in language symbols and abstractions. However, there is another language, composed of mathematical symbols and abstractions, which is known as statistics. Under certain conditions, statistical language can express the same thoughts more economically and allow us to reason in ways which the verbal approach cannot encompass.

For example, we may describe a fifth grade group of pupils in two ways, using the symbols and abstractions of verbal language and of statistical language.

VERBAL	**STATISTICAL**
There are twenty-eight pupils in this group—twelve girls and sixteen boys. They are of average intelligence, with no pupil extremely bright and no pupil extremely dull. Their achievement test scores average close to the national norms.	**N = 28** **G = 12** **B = 16** **IQ mean = 105; range = 96-110** **Stanford Achievement median = 5.3**

Figure 2. Verbal and Statistical Descriptions of a Fifth Grade Group

The illustration in Figure 2 makes use of *descriptive statistics.* In most instances, this kind of statistics enables us to work and think at a relatively high level of mathematical abstraction, unencumbered with the very weight, bulk, and inexactness of words. Another branch of statistics allows us to do still other things with our thoughts which neither a verbal approach nor descriptive statistics can encompass. These other things are associated with statistical tests of hypotheses and statistical inferences. This branch is known as *inferential statistics* and is based on the mathematics of probability.

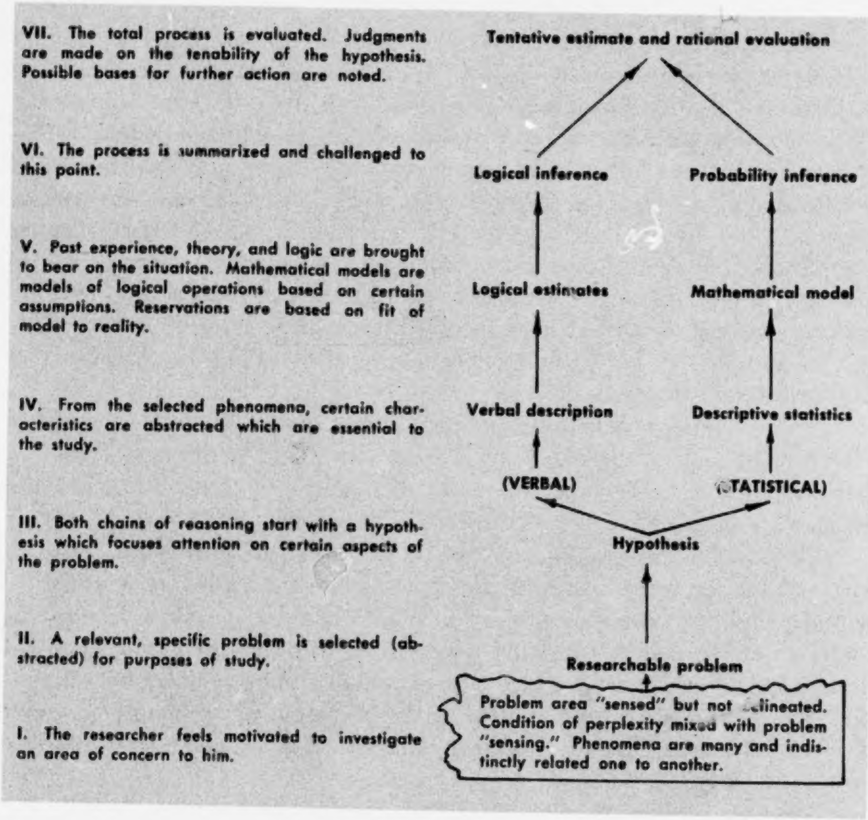
We know, for example, that any two measurements of phenomena (test scores) will differ to some extent just by chance or because of errors in the measuring instrument. Unless we have some way to account for the operations of chance or the errors, we have no means of estimating whether the difference we note between the two measurements is due to chance factors or whether it is significant. Probability tests give us a way to estimate the intrusion of chance balanced against the effect of our predictions. For this purpose, we choose an appropriate statistical test (mathematical model) to apply in the analysis of our data and then consult a corresponding table of probabilities, found in inferential statistics books. Such a table allows us to estimate the extent to which sheer chance, sampling errors, or measurement errors could or could not have produced the results we observe. Then we are able to report our findings as "significant" or "not significant" at a specified level of confidence (traditionally, at the one per cent or the five per cent level) in judging the tenability of our research hypothesis.

The foregoing discussion is a gross description of the procedures generally employed in following the chain of statistical reasoning. Later we will return to a more specific consideration of certain statistical tests and techniques which are especially useful for the educational practitioner.

Concurrent use of logical-verbal and statistical reasoning. Statistical analysis can never be a substitute for thinking. No amount of statistical sophistication can compensate for illogical conclusions. Most frequently, illogical conclusions result from errors of logic rather than from errors of technique. What is needed is a proper understanding of the contributions that both the logical-verbal and the statistical chains of reasoning can make to research efforts. Figure 3 portrays the interdependence and similarities of the two ways of reasoning.

Statistical methods may be regarded as methods for measuring, describing, and analyzing observations which have been made. But while the chain of statistical reasoning can lead the experimenter to greater precision in his attempt to quantify his observations, and while it can help to prevent him from drawing unwarranted conclusions concerning his data, this is not the whole job. Rules of evidence for drawing valid conclusions are also required. The researcher is still faced with the necessity for logical-verbal reasoning to interpret the meaning of his findings. As a result, there is a constant interplay between the statistical and the logical-verbal chains of reasoning in the research process.

Sometimes the experimenter can apply only logical-verbal methods. There are many problems and questions related to teaching and school management for which a statistical model simply will not suffice. It may be that ways to measure and express certain phenomena are not known, or that certain behaviors cannot be quantified satisfactorily. In such cases, the only available alternative is to proceed toward problem solution on



VII. The total process is evaluated. Judgments are made on the tenability of the hypothesis. Possible bases for further action are noted.

VI. The process is summarized and challenged to this point.

V. Past experience, theory, and logic are brought to bear on the situation. Mathematical models are models of logical operations based on certain assumptions. Reservations are based on fit of model to reality.

IV. From the selected phenomena, certain characteristics are abstracted which are essential to the study.

III. Both chains of reasoning start with a hypothesis which focuses attention on certain aspects of the problem.

II. A relevant, specific problem is selected (abstracted) for purposes of study.

I. The researcher feels motivated to investigate an area of concern to him.

Problem area "sensed" but not delineated. Condition of perplexity mixed with problem "sensing." Phenomena are many and indistinctly related one to another.

Tentative estimate and rational evaluation
Logical inference
Probability inference
Logical estimates
Mathematical model
Verbal description
Descriptive statistics
(VERBAL)
(STATISTICAL)
Hypothesis
Researchable problem

Figure 3. The Research Process—Two Chains of Reasoning[5] (*Read from bottom up*)

the basis of logical reasoning. But the experimenter can employ the same sort of careful thought patterns common to the more precise statistical forms of reasoning.

SEVEN STEPS IN THE RESEARCH PROCESS

Research always starts from a question or a problem of some sort: Does ability grouping produce more positive attitudes toward learning than does mixed grouping? Is reading comprehension improved by using an individualized approach as opposed to the traditional reading group approach? Should special classes or special schools be organized for gifted children? Do students learn more and better in specialized areas such as art, music, and physical education from specialists than they do from general teachers in self-contained classrooms? What is the relationship between oral language skills and abilities and written language skills? Does teacher attitude toward fringe responsibilities—playground and lunchroom duty, for example—influence student behavior in regard to such activities?

The purpose of research is to discover answers to questions through the application of scientific processes. To be answerable by research, questions must be asked in such a way that experimentation or observation in the real world will yield the needed information. They must be stated so that it is possible to apply scientific processes in seeking answers. The processes that are used have been developed in order to increase the relevance of the information gathered to the question asked. Scientific questions, of course, are not haphazard or sheerly speculative. They are prompted by reflections on the relationships of different forces and influences as we know them or might know them. Research processes allow us to test our hunches about how certain forces and influences affect or might affect one another.

As the various sciences have matured, they have developed efficient methods to communicate and discuss the methods of the research process. To facilitate discussion of these methods, it is convenient to organize the research process into a series of steps. Customarily, such organizations seem to carry the implication that step 4 always should follow step 3 and step 5 always should precede step 6. This unfortunate and mechanical interpretation of the research process is simply the result of confusing the logical form used in *talking about* research with the actual occurrences that go on in *doing* it.

There is no special merit in being able merely to name or identify a research method used. What is important is being able to devise a technique of study which will accurately solve the problem. No method or step, however ingenious, can be called effective unless it leads to well-founded answers. And it is the researcher's privilege to devise or invent procedures which will lead him as surely as possible to relevant concepts and facts. Any technique or method used should not be regarded as an end in itself but rather as the means to an end.

Acknowledging the artificiality of discussing the research process in terms of discrete steps, it seems useful to do just that. And since exposition is an efficient and widely used technique for communicating ideas, it will be economical to describe forthrightly the inner workings of the research process. The reader is warned, however, against the danger of accepting the descriptions of research as substitutes for direct observation and personal experimentation. If the descriptions serve as motivations to personal experiences that will exemplify and modify them, however, good may result.

Most descriptions of the research process eventually result in a series of five, six, seven, or more steps. For our purposes in administration and teaching, seven related and crucial steps have been identified.

STEP 1. Selecting the problem: Research is to the scientist what composing is to the musician, painting to the artist, and writing to the novelist. The artistry of the researcher is reflected in part in his selection of problems to be solved. Problems do not exist outside of persons. A dilemma does not become a problem until a person conceptualizes it. When the researcher selects a problem as worthwhile to act upon, he really selects his own conceptual way of looking at a portion of the world around him. When the two fit—the hypothesis and the real world—the hypothesis is taken to be usable with such events.

STEP 2. Accumulating pertinent knowledge and information: The custom of searching the literature bespeaks a preference for making an original contribution to knowledge and a correlative desire to avoid discovering things which are already known.

STEP 3. Latent period: The process of formulating ideas and schemes seems to require a period of intellectual gestation—a time when the researcher deliberately turns his conscious attention to something else. Many persons in such a situation find it advisable to sleep on the matter. Often, they awake to find that things have wonderfully straightened themselves out.

STEP 4. Idea or hypothesis formation: Two types of hypotheses are used in research: the statistical or null hypothesis (H_0), which is sometimes called the "hypothesis of no difference," and the alternative or research hypothesis (H_1), which can be accepted if the null hypothesis can be rejected. H_1 is a statement or proposition which is assumed to formulate solutions to the problem or problems.

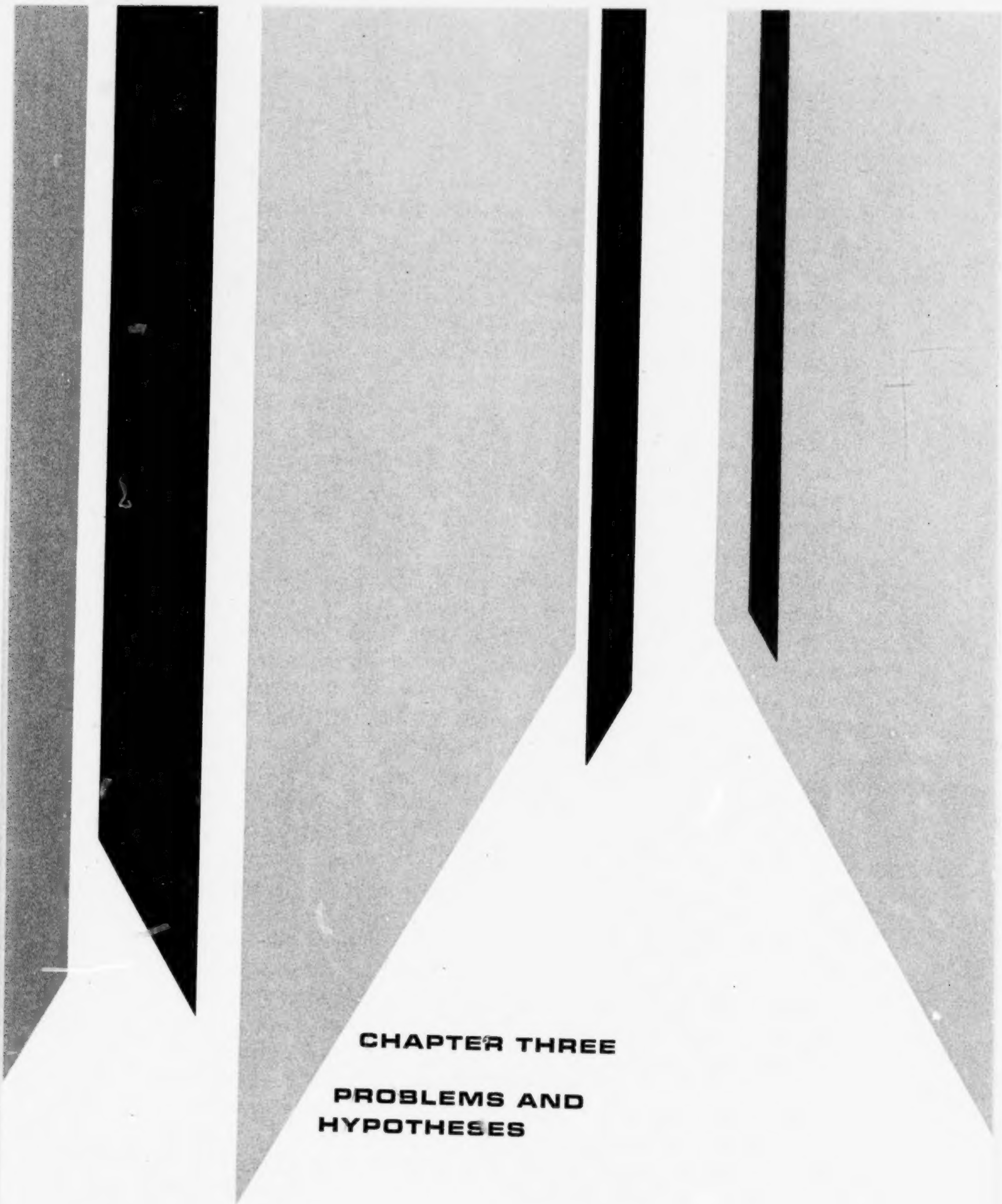
STEP 5. Designing the test of the hypotheses: The central outline of an experiment is simple: An "experimental" group is exposed to the experimental (independent) variable, while a "control" group is not; the two groups are then compared in terms of the observed effect.

STEP 6. Critical analysis and evaluation of observations made: The experimenter, to realize the maximum amount of information revealed through his data, must analyze logically, as well as statistically, the meanings of his findings. If "significant," what do the findings imply for the original problem? If not significant, how might the problem be approached from another direction?

STEP 7. Rejection or acceptance: In a statistical study, we always test the null hypothesis to find whether it can be rejected in favor of the alternative hypothesis. If we are willing to run the risk of being wrong in this rejecting, say 5 chances in 100, and our probability level is shown by the appropriate table to be .05 or less, then we may decide to reject the null hypothesis and accept the alternative (or research) hypothesis. But if our probability level is shown to be more than .05, then we cannot reject the null hypothesis and we conclude that the results obtained did not indicate a "significant" difference between the experimental and control groups.

In practice, the seven steps of the research process seldom follow one another in a fixed order. On the contrary, each step in the process does something to perfect the formation of ideas and to promote the location and definition of the problem. Each improvement in idea formation and problem definition leads to new observations that yield new data which aid in judging more accurately the relevancy of data already at hand. The elaboration of the hypothesis does not wait until the problem has been defined; it may come at any intermediate time. The introduction of any new step need not be final; it may be tentative and exploratory, later to be withdrawn or revised. In practice, two or more of the steps may telescope; some of them may be passed over hurriedly, and one or two may be chiefly responsible for the conclusions reached. The way the steps are managed depends upon the nature of the problem and upon the sensitiveness and imagination of the investigator.

One more idea should be noted about the handling of the research process. Frequently, steps are taken and observations are made that prove to be irrelevant to the main point being investigated. False clues are followed; fruitless ideas are entertained; superfluous techniques are tried. Because the solution to the problem is not known, it has to be groped for in the dark; lines of inquiry are started that in the end must be given up. The logical forms that characterize conclusions reached and accepted cannot, therefore, prescribe the way in which answers are sought when a condition of doubt and inquiry still exists. The printed reports of completed research projects present very misleading accounts of the ways research projects are carried out *before ideas become definite*. Seldom, if ever, do published reports tell of false starts, wasted hours, personal anguish, or personal exhilaration over unexpected successes. Yet all of these experiences and emotions are common to research, so long as it is in process.



CHAPTER THREE

PROBLEMS AND HYPOTHESES

In all of the sciences, research is basically a highly personal adventure. No research project is likely to be better than the creativity and imagination of the person who conceived it. This is another way of saying that technology cannot be a substitute for thinking. And thinking always occurs in a human cranium. Research problems always grow out of human thought. They may arise because the investigator is dissatisfied with the current state of affairs and wonders what would happen *if*. . . . They may arise from a feeling of discomfort prompted through simply not knowing enough about human behavior—which in itself is part of the very subject matter of education. Or research problems may grow out of the organized skepticism which is a necessary part of the scientific attitude. In effect, then, research problems arise from one of two general sources: the desire to know or understand for the satisfaction of knowing or understanding, and the desire to know in order to do something better or more efficiently.

Whatever its source, a problem is researchable only when research procedures can provide answers to it. One of the very necessary requirements in conducting research is the ability to recognize what is and is not researchable and to transform researchable problems into hypotheses.

RESEARCHABLE PROBLEMS

Perhaps one of the most helpful ways to illustrate what is meant by researchability is to consider first the kinds of problems and questions which are definitely *not* researchable. There are three types of questions commonly asked that fall into this category.

The first type of question that is not researchable is one that is so general and grandiose that a bushel full of findings—even if they could be secured—would seem infinitesimal in relation to the scope of the problem. The honest question, "Is it better to teach elementary school arithmetic by the 'meaning theory' than by more traditional approaches?" is one of these general questions with a thousand parts.

The second category consists of questions for which there might be answers but which are not researchable because adequate procedures for gathering the relevant information are not known. An example of this kind of question might be, "Does intellectual training in the basic subjects stimulate or depress creative thinking?" There may be an answer to this question, but procedures to identify creative thinking have just lately been experimentally derived and are not generally available for use in measurement studies. We can measure IQ by using any one of a dozen convenient standardized tests—but "creativity," no!

The third type of question which is not researchable deals with a choice or decision which involves values, as well as information. "Should the state legislature establish a compulsory system for the fluoridation of drinking water to improve children's dental health?" is a question of this type. The answer involves not only factual information, such as the effect of fluorine on the prevention of dental caries, but also religious beliefs such as those held by members of the Christian Science faith and political attitudes which tend to reject any governmental coercion. There is no scientific method for testing the *validity* of these values. Only the consequences may be known. Value judgments—as they affect the behavior of others and of the researcher himself—are a legitimate concern of educational researchers, and good techniques are known for obtaining information about them. The central position of value judgments in educational study lies in the fact that they are the formalized expressions of beliefs and feelings that impel human beings to action. But research cannot indicate clearly how people *should* feel in any one instance.

Manageable size. What does make a problem researchable? The first injunction might well deal with the importance of manageable size. The principle of parsimony enters the scene here. Specifically, this means that the question is better phrased if it does not lead the investigator into excessive expenditures of time, resources, and commitments.

Many questions profit from being reduced or from being divided into a number of subquestions to be dealt with in separate studies. For example, a question might be asked regarding the efficacy of teaching by means of closed-circuit television as compared with usual teaching methods. Matters of concern might include: 1) academic achievement of students, 2) attitudes of students, 3) opinions of teachers, 4) opinions of parents, and 5) administrative feasibility. Each of these subquestions could make a rather complete research question by itself, while investigation of all five might be necessary to secure adequate information in response to the original, generalized question which led to the investigation. The point is that a general question, broken down into its manageable parts, can be reformulated and recombined in order to find the most economical ways to secure answers.

Concreteness. The second requirement for researchability is that the question must be concrete and explicit. Concepts should be operationally defined, if possible. *What* is being questioned—attitudes, performance, a teaching method, certain concepts? *Who* is involved in the question—eighth grade students at Lincoln School, certain eighth grade students selected as a random sample, all teachers of the fourth grade in Farmer City, all parents of Miss Miller's seventh grade?

The merits of being explicit can be illustrated with two corresponding statements of a research problem. The question could be stated, "Might the first grade teachers of Jasper School, immediately following the first few weeks of acquaintance with the children in their rooms, be able judgmentally to list their children in order of readiness to read as accurately as can be done through use of the Metropolitan Readiness Test?" Certainly, this query awakens more definite mental pictures than the corresponding question, "Are teachers able to predict with some accuracy the potential achievements of students?" The first of these two questions is obviously more limited. But it is concrete and specific; it implies the kind of test needed; and it is apparently doable. The second question avoids mentioning *which* teachers, *which* students, *what* achievements, and under *what* conditions. Its boundaries are vague, and its empirical referents are lacking.

Measurability. A third qualification for a researchable question is that it should clearly define what is to be measured and should be related to adequate methods of measurement. This is simply to say that the researcher who does not know what techniques are available to test the implications of his question is in a poor position to formulate usable questions. The preceding question on readiness, for example, identifies two ways to measure readiness—teacher judgment and the Metropolitan Readiness Test. It might be that the accuracy of these two ways of measuring readiness could be determined through comparing the results they give with later school marks or achievement test scores. How well does year-end achievement correlate with beginning-of-the-year predictions of achievement as obtained from both teacher judgments and Metropolitan test scores? These additional measurement techniques suggest themselves almost spontaneously.

The researchable question gives clear and unmistakable clues to specific meanings: *who, what, how, when, where,* etc. An omnibus question is likely to beget a puzzled answer. And a question that lacks means for securing an answer is certain to beget frustration.

DEVELOPMENT OF HYPOTHESES

In research, problems and questions soon become transformed into hypotheses. A hypothesis is a statement or proposition which is assumed to formulate solutions to a problem. Whether the tentative proposition *is* the solution to the difficulty which originated the problem is the task of the inquiry.

The role of the hypothesis in scientific research is to suggest explanations for certain facts and to guide the investigations that will verify or refute the assumed explanations and lead to the discovery of others. If the original problem is carefully defined, the definition itself will suggest the kind of solution that is needed. When investigations of a problem are conducted purely at random, a multitude of facts will be turned up, but they will be so unrelated that their very number will add to the difficulty of the problem. It is quite possible for a research project to be swamped by the mere multiplicity and diversity of facts. The search for evidence is best conducted when the *most probable* meaning or supposition is used as a guide in exploring facts—especially facts that would lead to one conclusion and exclude others.

To describe the genesis of a hypothesis, we will illustrate with a paraphrased version of John Dewey's example of hypothesis making which appeared in his *How We Think*, published by D. C. Heath in 1933. Consider the familiar case of the automobile that perversely does not operate properly. Something is wrong, to be sure. But how to correct it cannot be decided until *what* is wrong becomes apparent. The untrained person is likely to make a wild guess or supposition and proceed directly in random fashion to poke here and hammer there, hoping to hit upon the right thing. The trained person will proceed in a different way. He will observe with great care, ruling out certain possibilities and narrowing the range of what might be the difficulty. He will use his knowledge of the structure of the machine to formulate the problem and detect the trouble. He will proceed in a distinctly nonrandom manner. Finally, he will select his best guess to act upon, but he will not rule out further possibilities by assuming that his best guess is certainly true. He regards his guess as a guide, as a working hypothesis, and uses it to check some facts, estimate others, and gather new ones. He may think that if the trouble is caused by insufficient fuel, and all other requirements for providing fuel to the motor are satisfied, *then* most likely the fuel pump is providing inadequate pressure, and he will look to see if *just* these conditions are present. His sense of the problem becomes more adequate; the supposition ceases to be a mere hunch and becomes instead a reasoned possibility.

The hypothesis implicitly states the problem and indicates the kind of solution that is needed. It forms an intellectual bridge between the problems and the tests of the most likely solutions to the problem. The hypothesis conditions the design of the tests, the collection of data, and the generalizations that may be drawn. Indeed, it is the pivotal point on which the whole research project turns.

Research hypotheses and null hypotheses. Hypotheses are frequently worded in "if-then" phraseology. A teacher may say, "If students choose the desks at which they wish to sit in the classroom, then they will have increased feelings of being accepted by the group." This teacher may derive his hypothesis from theories of social psychology or from the findings of other studies, such as J. L. Moreno's research into sociometric methods of grouping and regrouping. In such a case, the theory or the study may suggest the specific kind of data needed (sociometric choices); the particular information to be sought (increase or decrease in the number of mutual choices or isolates); and the design of the test (pre-test of sociometric choices, then honest follow-through of regrouping, and a post-test at some later time). But "if-then" phraseology is not the only way to state hypotheses. Any other assertion is acceptable as long as it contains a prediction that if certain conditions are provided, certain results can be associated with the conditions.

The kind of hypothesis we have been discussing is properly known as a *research hypothesis.* But if the study is statistical in nature, an additional kind of hypothesis is needed. This is known as the *null hypothesis* or the statistical hypothesis. The research hypothesis, as we have presented it, cannot be analyzed by statistical methods; the null hypothesis can be.

The null hypothesis is sometimes known as a hypothesis of no differences. For example, consider the question stated earlier about whether the first grade teachers of Jasper School can predict reading readiness as accurately as can be done through use of the Metropolitan Readiness Test. A statistical study might be planned to test the comparative accuracy of the Metropolitan Readiness Test and teacher judgment. The *research hypothesis* could predict: "Scores earned on the Metropolitan Readiness Test will predict readiness to read more accurately than can the first grade teachers of Jasper School through listing their pupils from most to least ready following the first month of school in the autumn." The corresponding *null hypothesis* would be: "There will be no difference between rankings of pupils in order of readiness to read as determined by the Metropolitan Readiness Test and by teacher listing." The null hypothesis is usually formulated for the express purpose of being rejected. If it is rejected, the research hypothesis may be accepted.

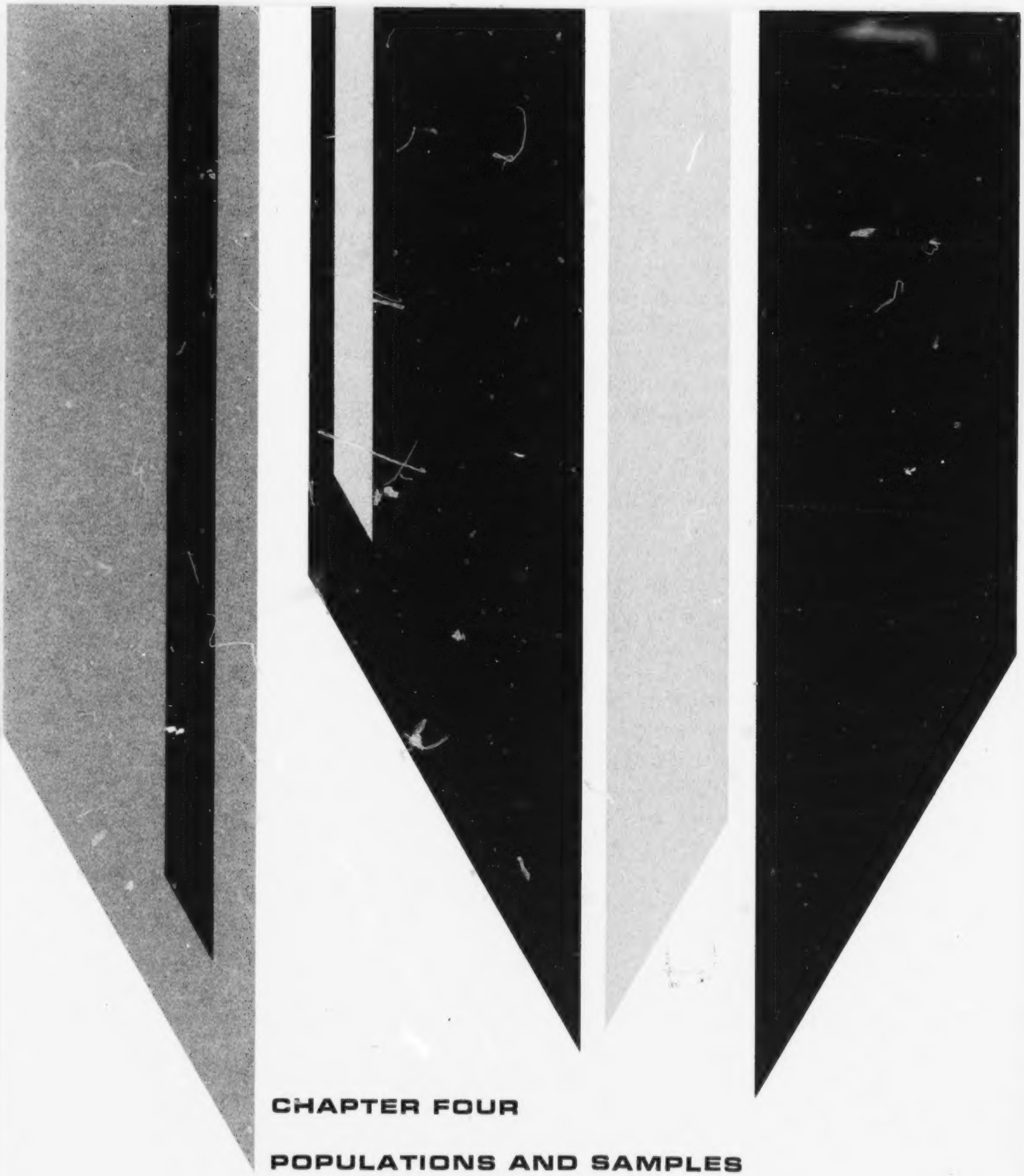
The symbol used for the research hypothesis is H_1. The symbol used for the null hypothesis is H_0. When we want to make a decision about differences, we test H_0 against H_1. In this case, if we let X stand for the

Metropolitan Readiness Test rankings and let Y stand for teacher judgmental rankings, we can express the two hypotheses in operational form: H_0 would be $X = Y$; H_1 would be $X > Y$ (X is greater than Y).

The nature of the research hypothesis determines how H_1 should be stated. If the research hypothesis simply states that the two measurements will differ, then H_1 is that $X \neq Y$ (X is not equal to Y). But if H_1 predicts the *direction* of the difference, i.e., that one specified group will have larger or more accurate scores than the other, then H_1 may be either that $X > Y$ (X is greater than Y), or that $X < Y$ (X is less than Y).

In testing the assertion that the Metropolitan Readiness Test will predict readiness to read more accurately than will teacher judgments ($X > Y$) —as measured by year-end achievement tests, for example—we cannot accept H_1 unless we are able to reject H_0. That is, unless we can reject the hypothesis that predicts "no difference," we cannot accept the hypothesis that predicts a statistical advantage for the Metropolitan Readiness Test scores. If we fail to reject H_0, then we must report our findings as those of "no difference."

The decision to reject or not to reject a hypothesis is a function of statistical probability. The mathematics and uses of probability will be considered later in some detail. For the present, we can be content with this much consideration of problems and hypotheses.



CHAPTER FOUR

POPULATIONS AND SAMPLES

The researcher becomes involved with populations and samples when he sets out to test a statistical hypothesis. A statistical question always relates to a group of individuals, rather than to a single individual. It asks what is true of the group. Statistical inquiries are of two types. One type of inquiry calls only for a description of the group of individuals actually observed. Summary measures, per cents, averages, or expressions of variability are computed from the observations. These measures comprise what is known as descriptive statistics.

A second type of statistical inquiry is more concerned with scientific investigation and utilizes inferential statistics. It involves a search for ideas and principles which have some degree of generality. If an investigation deals with matters of broad interest, the findings are usually applied to a much larger domain than the cases actually observed. This larger domain is called the *population*, and the group of cases observed is called the *sample*. The statistics obtained from the sample, which express certain characteristics of that group, are used to obtain information concerning the unknown group characteristics of the population. Such generalization from sample to population is known as *statistical inference*.

Statistical inference is concerned with two types of problems: estimation of population parameters and tests of hypotheses. While our primary concern in this book is with tests of hypotheses, we still need a clear understanding of population and sample.

In everyday language, the term *population* is used to refer to groups of people. We speak of the population of Chicago, or of the state of Utah, or of the United States, meaning by this all of the people who occupy a definite geographical space at a particular time. The statistician, however, employs this term in a more general sense to refer not only to defined groups of people but also to objects, materials, and measurements and to "things" or "happenings" of any kind.

For example, we may be interested in children with IQ's below 90 who belong to a high socio-economic group, or we may be interested in children with IQ's above 120 who belong to a low socio-economic group. If questions are raised about the proportion of children in a particular population or subpopulation with IQ's above or below a specified value, or if questions are raised about the relationship between IQ and socio-economic level, then these are questions of a statistical nature which deal with specified populations.

A further distinction may be made between *infinite* and *finite* populations. The limitless number of times a die could be rolled, or all of the variations that could be found among Arctic mosquitoes are examples of infinite or indefinitely large populations. And there are many populations which are actually definite or finite but which are so large that for all practical purposes they can best be regarded as infinite. The 180 million or so people living in the United States constitute a large but finite population which for most statistical purposes may be assumed to be infinite. This would not be true of a deck of cards, which may be thought of as a small finite population of 52 members. Nor would it be true of the twenty-three children in Miss Butler's fifth grade class, if they were defined as a finite population for a particular purpose.

If we had a research question concerning Miss Butler's fifth grade, we probably would prefer to study the whole small (finite) population, and we would not have to draw a smaller sample from it. But of course, we would be able to make statements only about this particular class and not about fifth grade classes in general.

Miss Butler's fifth grade, however, might be considered as a sample of all fifth grades in her particular school system. The extent to which we wish to generalize our findings from sample to population plays a major role in the procedures used. It may be assumed that any group of fifth graders is more like other groups of fifth graders than, let's say, like second graders or twelfth graders. Miss Butler might consider this similarity among fifth grade groups as a rough approximation of a sample drawn from a larger population. Subsequently, the accuracy of this assumption can be challenged against test data averages drawn from many fifth grade groups in the same school system. Miss Butler can use statistical tests of distribution very handily and profit from the knowledge which can be obtained. In addition, she can generalize her knowledge considerably beyond her particular class group. She can ask a question to determine whether and to what extent she would be justified in considering any particular fifth grade class assigned to her as a sample of all fifth grades she has taught in the past or might teach in the future. She could quite easily test the hypothesis that succeeding fifth grade groups in the school where she teaches will be significantly similar in crucial educational characteristics by applying the chi square test of the significance of differences. Of course, this would result in a longitudinal study stretching over as many years as she wanted to go, but it could be done provided both the foresight and patience were there.

But if for some good reason it became necessary to arrive at general statements concerning all fifth graders in the State of New York, then we would have a much larger problem. Although the population would be finite, for the sake of economy and reason, it should be treated as infinite. We most likely would want to select a representative sample of fifth graders or fifth grades, measure them in relation to crucial concerns, and infer back to the total population.

SELECTING A REPRESENTATIVE SAMPLE

There are three major steps in selecting a representative sample: 1) determining the nature of the population; 2) selecting a statistical model for analysis of the data; and 3) choosing a method for selecting the sample to make it as representative as possible. We will now briefly consider each of these steps.

Determining the nature of the population. In order to determine the nature or shape of a population, we generally rely on a body of knowledge from mathematics. We know that when enough measurements are made of any natural phenomenon—the size of oranges, weight of cats, height of adult human beings, IQ scores—and are charted on a graph, they tend to form a bell- or sombrero-shaped curve. This curve (or distribution) has come to be known as the *normal curve of probability* or, more simply, as the *normal curve.*

The normal curve charts the dispersion and variation which can be expected in any "nonselected" population of measurements, where chance alone is operating. It represents what will be the distribution of measurements or observations due to the probability of occurrences. It pictures the theoretical result of all chance distributions. In school parlance, the normal

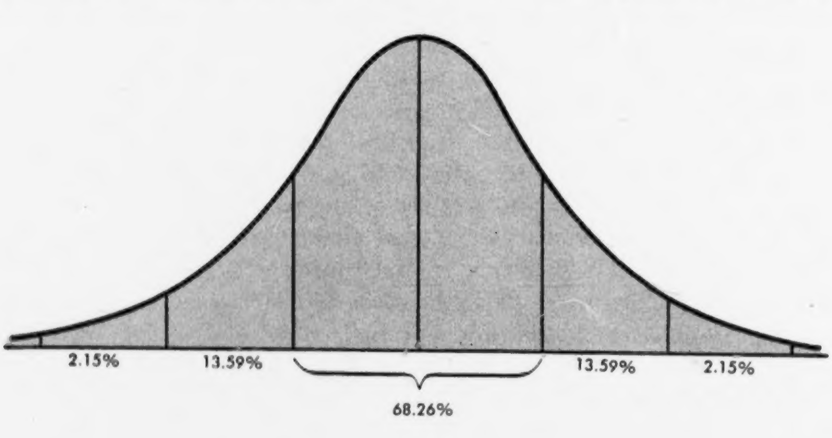


Figure 4. The Usual Pattern of the Normal Curve

curve is used to express the distribution of IQ scores, standardized achievement test scores, heights, weights, and numerous other familiar measurements. And it is worth noting that the curve tends to smooth out and to become more bell-shaped and symmetrical as the number of cases is increased.

Because of the steady recurrence of the normal curve when large numbers of phenomena are measured and charted, we are theoretically justified in assuming that large or infinite populations will also take the same distribution. Sometimes, in research studies, we are able to test this sort of assumption; more often, it is simply *assumed* to be true, without test. This practice of operating on the basis of reasoned, and not tested, assumptions is characteristic of statistical procedure, even when populations are not indefinitely large.

Selecting a statistical model for analysis of the data. We are faced with many of the same problems when we seek a statistical model or test for analyzing the data obtained from a sample. Customary parametric statistical tests of probability, such as the t or F distributions, are based on their own strong assumptions concerning the shape of the distribution which is assumed to characterize both the parent population and the sample. Prominent among these assumptions is that the observations are drawn from normally distributed populations: they must fit the normal, "chance" curve. The more numerous and stringent the assumptions, the more questionable the decisions arrived at through the use of any statistical test of probability. We know that many populations or samples are not distributed normally. What then? What happens when the measurement is not of sufficient strength, not truly numerical? What sort of statistical model might we select which is not hedged with strong assumptions about the shape of the distribution and measurement requirements?

There are satisfactory answers to these questions. Just as the parametric tests are based on distribution and measurement assumptions, so there is a branch of statistics known as distribution-free, or *nonparametric*, statistics. With these nonparametric models we may say, "Regardless of the shape of the distribution, we may observe that the probability of occurrence is . . . ."

However, we should note that the parametric tests with the strongest assumptions are also the most powerful tests. When those assumptions are valid, these tests will be most likely of all to reject H_0 when H_0 is false. But the assumptions and requirements underlying these more powerful tests must be met before confidence can be placed in any probability statement.

We will shortly return to the use of distribution-free statistical models when we consider the analysis of data obtained from populations and samples.

Choosing a method for selecting the sample. Sampling, of course, is a necessary procedure regardless of whether we are concerned with the shape of the parent population. The same procedures of sampling apply to all cases where the gross number in the population can validly be reduced to promote ease and convenience.

There are many ways to draw a smaller sample from a larger group. But in all the many different types of samples, the most important characteristic is *randomness*. By this is meant that every single sampling unit in the population has an equal chance of being drawn into the sample. In the most direct form, the sample might be drawn by simply writing on a card the name of each individual included in a finite population, scrambling the cards in a box, and then selecting, without looking, the number of names desired for the sample.

Something like this was done in selecting men to be called for military service during World War II under the National Selective Service Act. When each individual registered with his local draft board, he was assigned a number, ranging from one to several thousand. Each of these numbers was then encased in a capsule, the whole group of capsules thoroughly mixed, and all of them placed in a large fishbowl. The President of the United States, blindfolded, reached into the bowl and drew out one capsule after another. The local draft boards sent out induction notices in the order of the drawing to the men holding the selected numbers. Placing all the numbers in similar capsules, putting them in the same bowl, mixing them, and then blindfolding the drawer, made this type of sampling truly random and nonbiased. Such a procedure is known as *simple random sampling*. In using this procedure, every effort is made to reduce the possibility that selection of the sample will be influenced by either conscious or subconscious bias and to leave instead the decision solely to sheer chance.

Questions about the optimum size of the sample in research logically arise at this point. Statistical calculations can give us an idea of the degree of confidence which can be placed in the accuracy of a sample of a given size and of the sampling error which may be present. By and large, however, the accuracy of any sample can only be estimated. The rough rule of thumb is that small populations require a proportionally larger sample than do large populations. For example, it might be preferable to draw a 50 per cent or larger sample of the 100 sixth graders in a particular school, while a 10 per cent or smaller sample of the 74,500 sixth graders in a particular state might be adequate. In both cases, random selection would be necessary.

Sampling from populations of very large size can become inordinately time consuming and tedious. Easier and defensible results can be obtained by use of a table of random numbers such as may be found in most textbooks on statistics. These tables contain sets of numbers that after careful examination have shown no evidence of systematic order. Before using the table, it is necessary to number all the individuals in the population to be studied. This can be done starting with 1 and proceeding to 50, or 99, or 199, or 1999, or whatever figure equals the total population number. The table is then entered at some random starting point—perhaps by making a blind stab at the page with a pencil—and the cases whose numbers come up as one moves down the column of numbers are taken into the sample until the desired number of cases is obtained. The selection of any given case places no limits on what other cases can be selected from the population.

Other types of sampling are available, such as ***stratified random sampling***. This approach uses the same "blind but fair" techniques which characterize random sampling. But when a population is composed of several subpopulations—for example, Negro-white, culturally advanced-handicapped, or intellectually accelerated-retarded—it may be divided into two or more strata. A random sample is taken from each stratum, with the subsamples joined to form the total sample. There is no attempt to make stratified samples a replica of corresponding populations; the intent in stratified random sampling is only to take into account the anticipated homogeneity of the defined strata with respect to the characteristic which is being studied. And of course, each of the subsamples can be investigated separately.

SPECIAL PROBLEMS IN SCHOOLS

Knowledge of population and sampling techniques is valuable to the extent that it can be used feasibly and accurately. But the conditions surrounding schools, their organization, and the teaching that goes on within them set rather severe limits on choices of populations to study and on the kind of samples which can be drawn from them. Some attention to these imposed limits and restricted choices will complete our brief treatment of populations and samples.

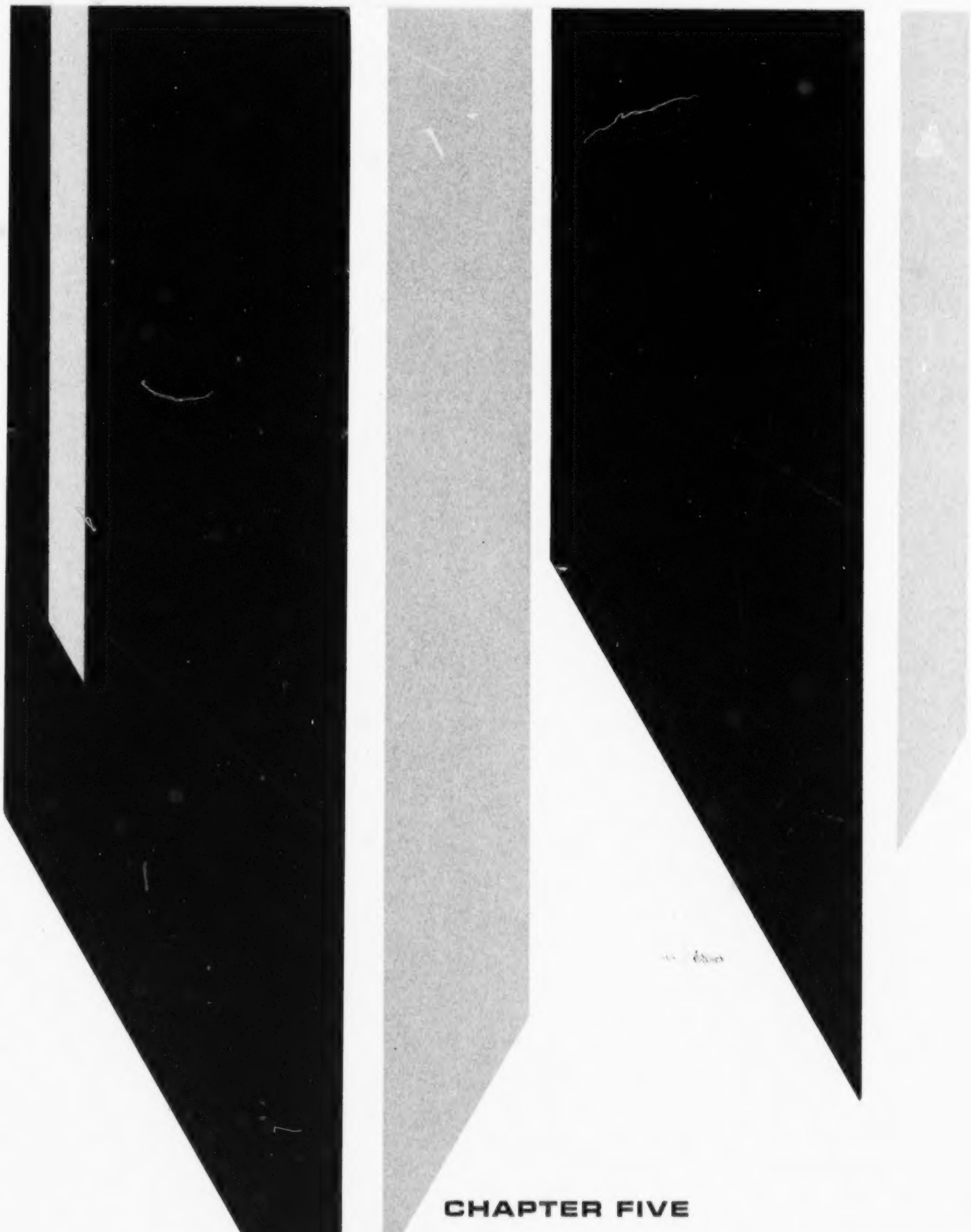
In the first place, the students in any single school or school system, and the teachers and administrators therein, are *not* examples of natural or unselected populations. They represent highly selected and particular kinds of captive groups. Furthermore, the longer children are in school, the keener becomes the quest for survival and the larger becomes the attenuation rate, thus reducing the number of certain types of children in regular attendance. Just witness the national concern over school dropout rates. More than this, the various ways in which children are grouped, once they enter schools, result in quite unnatural, unrandom aggregates of individuals. The truly great variations in grouping patterns employed among numbers of separate schools and the select influences on their populations greatly complicate the process of generalizing from sample to universe.

A further complication arises from the fact that schools are deliberate, goal-oriented social institutions, remarkably insulated from scientific manipulation. To a significant degree, researchers in education are limited in the extent to which they can experiment with the organization of grade groups, with instructional requirements and the allocation of content, and with other policies about the conduct of the school. In other words, researchers dealing with the schools—with deliberate, goal-oriented institutions—must to an important extent work with the groups that they already have. Because of the very nature of the school as an institution, they cannot select different groups out of a gross, chance population and begin afresh to organize an imaginative educational system composed of chance students.

Coincident with this observation, it is impossible to avoid thinking of the Amidon School in Washington, D.C. Here was a once-in-a-lifetime opportunity to choose the students for a school without a contiguous group of parents, house them in a new building, and manipulate the instructional program. Of course, this opportunity came about through certain miscalculations in urban redevelopment which resulted in a new school building without a community and without children to be educated. There was no other choice but more or less to manufacture a student population by importing children from other areas of the city. What a wonderful chance this would have been to undertake some careful, imaginative, scientific research!

In any event, the particular problems of population and sampling in relation to the schools may be responsible for the tendency of educational research to deal with the "samples" (class groups) which are already there. Such "samples" are assumed to be representative of a larger population of theoretical class groups, and findings are then generalized from samples to populations. And then teachers and administrators are sometimes surprised when research findings simply do not seem to transfer from the samples involved in the studies to other classrooms which make up the whole population!

There are ways to handle these problems of population and sampling which arise in educational research, and we will consider them later on when we get to a discussion of statistical analysis and to a consideration of research designs and decisions.



CHAPTER FIVE

VARIABLES AND PROBABILITY

The central outline of an experiment is simple. An experimental group is exposed to the experimental variable while a control group is not; the two groups are then compared in terms of the observed effect. It should be stressed at this point that some sort of a control group is essential in experimental research. However, practical limitations frequently demand modifications and inventions in establishing experimental controls.

We have just discussed some of the difficulties involved in securing adequate populations and samples. The attempt to provide reasonable similarity between experimental and control groups immediately doubles the difficulties. For the purposes and conditions of most social research, the "elegant" research design with "pure" experimental and control groups, nicely matched in all important respects, is seldom feasible. Rather than lose the potential knowledge to be gained through experimental research, we tend to deal with estimates or approximations about the meaning of our data. It is one thing to recognize what the ideal set of conditions surrounding research may be; meeting these conditions in the ordinary, ongoing practical world is something else and frequently calls for a good deal of imagination and inventiveness if the ideal is to be even approached. An understanding of variables and probability is important equipment for dealing with the realities that often complicate experimental research.

INDEPENDENT AND DEPENDENT VARIABLES

In research, there are two broad kinds of variables to be reckoned with. First, there is the *experimental variable* which is the crucial characteristic, occurrence, or procedure being investigated. This is also called the *independent variable.* Customarily, the experimental or independent variable is symbolized by the letter X. Second, there is the *dependent*—or *criterion* or *predicted*—variable which is the assumed effect of the independent variable. It is called the dependent variable because the prediction of it depends on the functional relationship or value of the independent variable. Customarily, the dependent or criterion variable is symbolized by the letter Y.

Thus, an experimental hypothesis asserts that a particular characteristic, occurrence, or effect (X) is one of the factors that can be associated with another characteristic or occurrence (Y). Or, the hypothesis may predict that if X, then Y. Studies designed to test such hypotheses must yield data which will lead to the inference that X can or cannot be associated with Y under specified conditions.

An experimental design provides both greater assurance and greater efficiency by making possible the simultaneous collection of several lines of evidence. More specifically, the use of experimental designs makes possible the collection of three major types of evidence necessary to testing hypotheses: 1) evidence of concomitant variation—that is, evidence that the independent variable (*X*) and the dependent variable (*Y*) are associated; 2) evidence that the dependent variable (*Y*) did not occur before the independent variable (*X*); and 3) evidence ruling out other factors as possible influences on the dependent variable.

Evidence of concomitant variation. Evidence of the first type is secured very easily in an experiment. The investigator knows which subjects have been exposed to the experimental variable (*X*) and which subjects have not. He measures all subjects in terms of the dependent variable (*Y*). He can then determine whether *Y* occurs more frequently among those subjects who have been exposed to *X* than among those who have not, or whatever specific relationship is predicted by his hypothesis.

Evidence that the dependent variable did not occur before the independent variable. The second type of evidence needed to test a hypothesis is secured in one or both of two ways. The investigator can set up the experimental and control groups in such a way that it is reasonable to assume that they did not differ in regard to the dependent variable before the experimental group was exposed to the independent variable. Or the investigator can measure both groups in relation to the dependent variable before one group is exposed to the independent variable. This results in a "before" measurement and an "after" measurement, or in a pre-test and post-test type of design. The differences between the pre- and the post-tests for both groups provide the desired comparison. The intent behind this requirement is not to demonstrate that one or both groups have none of the prior knowledge, or attitudes, or qualities being investigated, but to demonstrate that the experimental group did not have *more* of the factors being experimented with before the start of the experiment.

Evidence ruling out other factors. There are several ways to obtain evidence ruling out other factors as possible influences on the dependent variable. At least four major sources of uncontrolled factors may interfere with an experiment.

First, *the measurement process* used in the experiment may affect the outcome. If people feel that they are being used as guinea pigs or if they feel that they are being tested and must make a good impression, or if they feel that they are different from other people because they are defined as experimental, the measuring process itself may distort the results. The "before" measuring may crystallize attitudes related to the experimental variable. The "after" measuring may be influenced if the subjects reflect test fatigue or become bored or defensive. The subjects may try to keep their second responses consistent with their first responses or may try to vary them if they feel that variation is expected. If there is reason to suppose that any of these influences is present, alternative designs can be employed to lessen the impact of the measuring process itself.

Second, normal changes related to the *maturation and development* of the subjects may be confused with the results of experimental treatment. This kind of unplanned interference is especially critical whenever an experiment extends over a long period of time. However, if the maturational changes can be assumed to be approximately the same in the experimental and control groups, the effect of maturation can be ruled out by comparing the two groups. Sometimes, interim measurements during the course of a long-range experiment will indicate whether the two groups are becoming noticeably unlike.

Third, in any social experiment, *unplanned contemporaneous events* may affect the experimental outcome. If during the course of an experiment in the teaching of science, a foreign nation were to claim new and frightening scientific discoveries, attitudes toward the teaching and learning of science in American schools might be influenced—as indeed they were in the familiar instance of Sputnik I. If such an event were to affect the experimental and the control groups in the same way, no problem would be created since an effect common to the two groups could not be a cause of difference between them. However, it is very difficult to ascertain the relative impact of deep-seated, unplanned events on two separate groups. Such events cannot be controlled, either in the sense of holding them constant or in the sense of deliberately manipulating them. But some experimental designs have been devised to take account of this possibility, at least in part.

Fourth, *characteristics of the subjects* that are relatively enduring or are the result of past experience may introduce systematic differences between the experimental and control groups. In order to rule out such differences in relevant factors, the investigator may set up his experimental and control groups so that it is reasonable to assume that they do not differ widely in the distributions of age, sex, intelligence, social background, academic achievement, and other factors seen as relevant to the experiment. Or he may measure them before the experiment in terms of such factors; or he may do both. All of these factors relating to subjects are also "variables." But they are controlled variables, controlled to eliminate their influence on the dependent variable which is being studied. At times, groups rather than individuals are matched. For instance, it may be sufficient to make sure that an experimental and a control group are similar in relation to averages and ranges of age, intelligence, achievement, and social background, and any other factor presumed to introduce systematic differences. This practice is known as *frequency distribution control.*

STATISTICAL TESTS OF PROBABILITY

It may be that we have carefully allowed for the collection of these three types of evidence necessary to testing hypotheses, but there still remains one more type of evidence which must be obtained. Without it, we still might believe that we had established a clear relationship between *X* and *Y* when, in fact, we had not. It might easily be that *Y* occurs as we observe because it is an accidental or fortuitous occurrence. This fourth type of evidence is secured through statistical tests of probability.

Probability enables us to allow for the intrusion of sheer chance or sampling error. The extent to which chance might have entered into a research study is expressed as *significance.* If chance might have played a large part in the findings obtained, the significance value, expressed as a percentage, will be large; if chance played a small part, the significance value will be small.

The problem of deciding on the significance of research findings related to schools may be directly associated with the fact that almost any teaching method or organizational practice is bound to result in some sort of change in the students exposed to it. Most teachers and administrators know that it would be extremely difficult to prevent students from changing (learning) to some extent, regardless of the approach used. This is all part of the dynamic nature of human beings; they simply will not stay still. Even if a student is left alone, he will construct an activity out of something, if for no other reason than to have something to do. Because of this incessant momentum, there is always the real possibility that the gains in learning observed by the teacher may have resulted from sheer chance happenings rather than from the activities deliberately planned or assigned. And to complicate matters still more, the students of any class, in any learning activity, will show differences in quality, quantity, and speed of learning. Some may give considerable evidence of gains in the desired direction, and some may even show a loss. How then can we determine whether the results of any research project related to schools, children, and teachers can be attributed to chance or to the effect we predict?

AN ILLUSTRATION

Let's see how probability and tests of significance are used in experimental research. Take the case of Mr. Black, sixth grade teacher, who wanted to experiment with ways to increase the number of library books read by the students in his room. He felt that the students would read more books if they had the additional motivation and recognition that could come about through writing and publishing book reviews with a by-line in the class newspaper.

Mr. Black began by consulting his records to determine the number of books read per month by each student. Then he listed the students' names in order, from most to least books read per month. After planning with the class, he encouraged them to write reviews of the books they read and proceeded to compile another set of records. Two months later, he again listed the names in the same way he had done before to determine whether the predicted increase in reading had occurred.

In following these procedures, Mr. Black had used a pre-test and post-test, single-group design which utilized sixth grade students as their own controls. That is, the pre-test (number of books regularly read) served as the control group, and the post-test (number of books read following the writing of reviews) served as the experimental group. He felt confident that he had given adequate attention to concomitant variation, that he had evidence of the fact that increases in reading (Y) had not occurred before the writing of reviews (X), and that other factors had been controlled by presenting the project to the class in a calm manner as a part of regular, everyday classroom work.

But he still was faced with the necessity of accounting for the operations of pure chance. Mr. Black observed that of the twenty-eight students in the room, eighteen had increased their reading, three had read exactly the same number of books per month that they had read before, and seven had read fewer books. He could see that there had been a change for the better with the eighteen students. But what about the ten who showed either no change or a loss? Were these results significant, or could this distribution have been the result of sheer chance? Unless he could determine whether the results were simply chance findings, he could not decide whether he had made significant discoveries about ways to stimulate his class to read more library books.

There are very good, feasible methods for Mr. Black to use in determining the significance of his findings. He can turn to some of the probability tests in statistical inference, apply his observations (data), and determine whether the results are *statistically* significant. His aims are to judge the significance of the data and to make maximum use of the information gathered. These are the principal problems with which statistical methods are concerned. And Mr. Black, sixth grade teacher, can learn to use the statistical tests he needs in just a few weeks.

What "probability" means. It will be helpful to take a moment or two to look more closely at what Mr. Black would be doing if he applied an inferential test to his data. He would start with the realization that a statistical question always relates to a group of individuals rather than to single individuals and reveals information about the characteristics of the group. Next, he would know that the results he would derive from the test would take into account the probability of chance occurrences and enable him to say, with a specified degree of confidence, whether the differences he noted in the number of books read could legitimately be attributed to the writing and publishing of reviews in the class newspaper.

He would realize that the results of the statistical test would be expressed in terms of the probability that his decisions might be wrong. Even if he finds evidence to indicate that writing reviews is an effective stimulant to reading books, he must allow for probable error in deciding that this is so. There are very few facts in life about which we can be absolutely certain. We are far more justified in thinking in terms of probabilities than of certainties.

For example, insurance actuaries have calculated the risk of people in the United States dying at various ages. Their figures let us know that a twenty-year old person has one chance out of approximately 740 that he will not see his twenty-first year, and that a thirty-four year old person runs the risk of one chance in approximately 440 that he will not become thirty-five. However, if we were to make predictions that a certain number of any particular group of twenty-year olds or thirty-four year olds would die during those years, we would probably be wrong to a certain extent. Yet if our predictions were consistently more right than wrong, we might regard them as statistically significant.

But of course, Mr. Black would not be dealing with insurance tables. He would be using the mathematical materials and tables which have been worked out to express the probability of occurrence, such as the chance of drawing one of the four aces in a game of poker. (Parenthetically, we might make the observation that all statistical procedures originally grew out of the study of games of chance.) But to continue with the explanation of probability: If two or more events are equally likely or probable of occurrence, this fact can be represented by a ratio

—for example, 1/2 as in coin tossing, 1/6 as in throwing dice, or 1/52 as in drawing any particular card from a standard deck. The ratio means that if there are N possibilities, the probability of the occurrence of any one of these possibilities is $1/N$. Thus, if a coin is tossed, the probability of its falling heads up is 1/2, and the very same value holds for the probability of its falling tails up. And since $1/2 \times 2 = 1$, these two possibilities are mathematically and logically exhaustive. Similarly, if a die, which has six sides and thus six possible ways of landing, is thrown, the probability of any one of the six sides coming up is 1/6. Six times this fraction equals one, which again is logically exhaustive.

From this way of stating probabilities, all manner of more complex expectations or probabilities can be derived. For instance, the expectation of obtaining two heads from two tosses of a coin is no longer 1/2 but is $1/2 \times 1/2$ or 1/4; the expectation of obtaining three heads from three tosses is $1/2 \times 1/2 \times 1/2$ or 1/8. Or the expectation of rolling two sixes, or any other two numbers, on a pair of dice is $1/6 \times 1/6$ or 1/36. And it has been found that if all possible combinations of heads and tails on ten successive tosses of a coin or on the simultaneous toss of 10 coins are computed, the resulting distribution turns out to be the normal, bell-shaped curve.

What "significance" means. From the type of mathematical model just described, it is possible to say just how likely a given outcome is if chance alone is operating. If, therefore, in a situation where an independent variable has been deliberately introduced the results are such that they might easily be attributed to chance alone, we do not consider that anything has been demonstrated. But if a result is obtained which would be expected by chance only one time in a thousand ($P = .001$) then we can say that it would be very unlikely for chance, rather than the variable in question, to be influential. This is the logic of the null hypothesis—the hypothesis that the difference obtained between the experimental and control groups is *not* due to the deliberately introduced independent variable but is due instead to an accident or error of sampling. What the evidence we collect does is to discredit the null hypothesis—the hypothesis of no deliberately produced difference. It is here that the calculation of probability may help us to decide whether we have achieved evidence in which we can have reasonable confidence.

It is just this sort of calculation and decision that Mr Black would undertake in analyzing his data for significance. Fortunately, he has access to statistical formulas and tables which will simplify the computation for the various tests he might use. And he can use the newer distribution-free tests which have been made generally available since 1956. In using any of the tests, Mr. Black would have some decisions of his own to make. He would have to decide how much risk he is willing to take that he might be wrong. Would he be willing to accept his experimental results if he were wrong one time out of a hundred? Five times out of a hundred? Twenty times out of a hundred? Depending on the crucialness of his research question and the accuracy of his measuring instruments, how right might he insist on being, and how much risk is he willing to take that he might be wrong?

In the language of statistics, probability values range between 1 and 0. The probability that each of us will someday die is 1. The probability that any one of us will never die is 0. The probabilities that people will live to given ages may be expressed as a decimal fraction somewhere between 1 and 0. All probabilities may be expressed this way. For instance, if we toss a coin, we can state that the probability of obtaining a head is one out of two. We write this as $P = 1/2$, or .50.

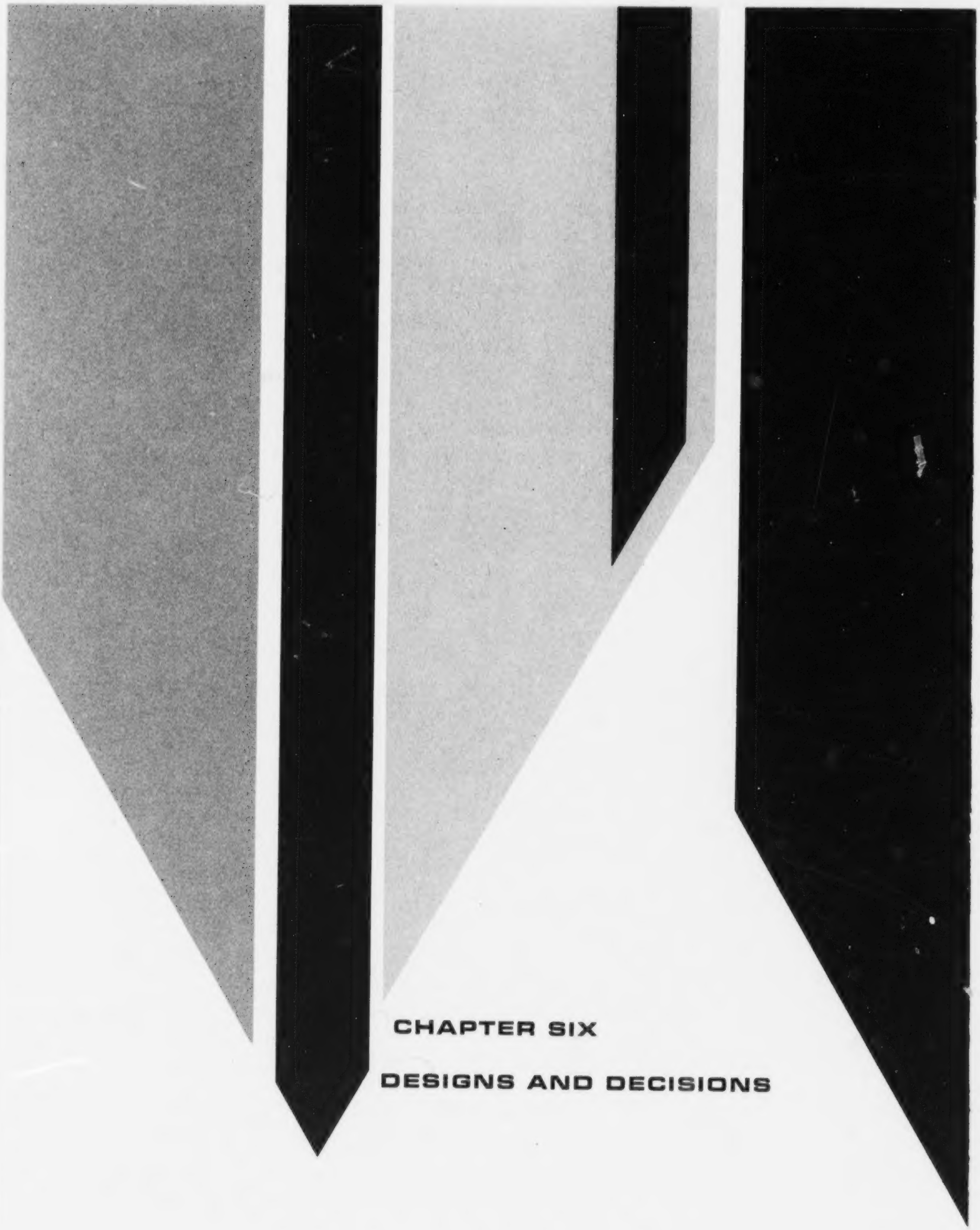
Mr. Black will state the risk he is willing to take this way. If he wishes to restrict the probability of being wrong to one chance out of a hundred, he will test his data at the ".01 level"; if he wishes to restrict it to five times out of a hundred, he will test his data at the ".05 level," and so forth. The statistical tables he will use are designed in terms of these alternative choices, and he can read directly from them the significance of his data at the level of risk he is willing to take. Of course, if he chooses to work at the .01 level, he will not be able to accept his experimental findings and consequently be able to reject his null hypothesis as frequently as he might be able to do if he worked at the .05 level. And he could possibly reject the null hypothesis many, many times more frequently if he worked at the .10 level. But the decision is Mr. Black's, and he should report his research findings as "significant at the ....% level." Then other people can decide whether, for their purposes, they wish to accept or reject his findings at the probability level he reports.

Working with these statistical methods, Mr. Black would know that significance is never a matter of true or false, of significant or not significant. He would know that significance is always expressed in levels of probability of occurrence in an infinite number of theoretical tries, stated as a percentage.

Incidentally, the table of probabilities for the distribution-free Sign Test shows that for the way Mr. Black's students were distributed in regard to increased reading of library books (18 read more, 3 read the same number, 7 read fewer books), the experiment was significant at the

.022 level. Sheer chance could not have accounted for these results more than 2+ times out of a hundred theoretical tries.

We should end our consideration of probability with a word of restraint. With a large enough number of cases, extremely small differences may be "statistically significant." For example, with a sufficiently large experimental and control group, a difference of one point between their mean or median IQ scores could make the difference "significant," although it is hard to see what importance such a small difference might have. It is sometimes difficult to recognize any real import in trivial findings. Statistical manipulation is no substitute for thinking.



CHAPTER SIX

DESIGNS AND DECISIONS

The various phases of the research process can be brought under control by designing the research. To design is to plan; that is, design is the process of making decisions before the situations arise which require the decisions to be implemented. It involves deliberate anticipation of situations which are expected to arise and an effort to bring them under control. In planning the design, it is safest to go right back to the origin of the inquiry and ask, "Why am I doing this particular thing? Will it really tell me what I want to know?" After all, it is rather poor policy to carry out an experiment without having a clear-cut idea in advance of just what is being tested and without an equally clear tentative plan for putting the test into effect.

The ideal experiment is sometimes described as one in which all relevant variables are held constant except the one under study, the effects of which are then observed. But this is not the best pattern in some cases, and in most actual situations encountered in the behavioral sciences, it is not practical. Consequently, some device is needed to help correct or account for the effect of unanticipated or uncontrolled variables during the course of an experiment. The introduction of *controls* provides this sort of device. In most cases, the real object of the experiment is some sort of comparison, and with the introduction of controls, this takes the form of a comparison between the experimental and the control groups. But in educational research, it is extremely difficult to secure control groups precisely matched with experimental groups on all relevant variables. Hence, it is important to have at hand a variety of ways to provide for some sort of control in experimentation. Having a choice among several research designs makes it possible to select some sensible pattern of study which will not sacrifice the very real advantage of controls simply because of the difficulties of securing them.

The entire design of an experiment has the function of providing for the collection of evidence in such a way that inferences of a relationship between the independent and dependent variables can be drawn as surely as possible. To achieve this, many research designs have been developed in the various sciences. We will now describe and illustrate five designs which are particularly usable in studies of teaching and administration.

DESIGN 1: "BEFORE-AFTER" STUDY WITH CONTROL GROUPS

This design is one form of the classical control group design. It uses two groups of individuals which have been reasonably matched on variables relevant to the experiment or which have been randomly assigned to the two groups. Usually, both groups are measured in relation to the dependent variable at the beginning (Y_1) and at the end (Y_2) of the experimental period. The experimental or independent variable (X) is introduced in the experimental group only; it is not introduced in the control group. Since the experimental and control groups are both pre-measured, subject to uncontrollable factors and vulnerable to unplanned, contemporaneous events, the differences between the scores of the two groups should approximate a direct indication of the experimental variable's influence alone.

The pre-measures of the experimental group (Y_1) and the control group (Y'_1) may be used simply to check the initial absence of difference between the groups. With this much assured, the final scores of the two groups (Y_2 and Y'_2) may be compared to analyze the influence of the experimental variable (X). Or we may use both the pre-measure scores and the post-measure scores of both groups to calculate *change scores.* Then we compare the change in the experimental group (d) with the change in the control group (d').

Example of a "before-after" design with one control group. Samuel A. Kirk's *Early Education of the Mentally Retarded* contains an example of a "before-after" experiment using one control group.[6] The experiment reported in this book was undertaken to determine the effects of pre-school education on the social and mental development of young, educable, mentally handicapped children. The report on this experiment has been selected for our purposes here because, in addition to illustrating the design, it also may illustrate the kind of modification in the "ideal" design pattern made necessary by restrictions encountered in actual situations.

To quote directly from the published report of this research study:

"The difficulties with experimentation in this area are numerous. Many factors are beyond the control of an experimenter when he is dealing with children, parents, and communities, whose movements and operations are their own affair. It cannot be managed as simply or as accurately as experiments with rats or human experiments using nonsense syllables. One has to be satisfied with an approximation of an experimental design. The aim of the present project was to set up an ideal experimental design and approximate it as closely as possible.

"Under ideal conditions one would obtain a group of children within a specified range of age and mental ability, randomize a control and experimental group, subject one group to training, and leave the other group at home. Identical tests and observations would be made on both groups before and after the training period. After five years, comparisons would be made between the two groups. No children would get sick, no families move out of town, and all would attend or not attend as directed. Since this ideal is impossible, an approximation of it was carried out, and deviations in terms of loss of cases, randomization, and so forth, are explained in later pages. To obtain experimental evidence on this problem it was necessary to set up a project as follows:

"1. Community Experimental Group. This experimental group consisted of 28 preschool mentally handicapped children from the community who were given an enriched nursery school environment from nine to three o'clock each day. At about six years of age these children were placed in a regular first grade or a special class. There were about 15 children at any one time in the nursery school.

"2. Community Contrast Group. This group of children consisted of 26 young mentally handicapped children found in the same community or in neighboring communities. They were examined at the same intervals as those in the experimental group, but they did not attend a preschool. The term *contrast group* is used here instead of *control group*, since these children and the children in the Community Experimental Group were not randomized in the usual sense. It was impossible to find sufficient children at any one time to randomize them."[7]

In this study, Kirk also used the same design to conduct a parallel study of feebleminded children who had been committed to state institutions. This concurrent, but separate, study was set up as a replication of the community pre-school experiment. The two institutional groups being compared were named the Institution Experimental Group (15 children) and the Institution Contrast Group (12 children). The purpose of conducting this concurrent replication was twofold. It served as a check on procedures employed in the community experiment, and it revealed information concerning the effects of specialized training on the development of institutionalized children.

The control group-experimental group pattern is admirably suited to experimental studies, in fact, its basic design elements are essential in some form if we are to have a firm basis for inferences about the effect of the experimental variable. But the full design is beset with difficulties in practical situations. The investigator should be aware of the potential difficulties if he is to make allowances for them, as was done in the Kirk study.

Advantages and limitations. As we have pointed out, this design allows us to take account of the effects of initial measurement, uncontrollable factors, and contemporaneous events in determining the influence of the experimental variable. The use of a control group allows us to avoid confusing experimental results with irrelevancies. We can relate the two groups to each other, control or manipulate the relevant variables, and compare the experimental variable group with the nonexperimental variable group. If the two groups have been satisfactorily matched before introduction of the experimental (independent) variable to the experimental group, and if adequate control is achieved to keep experimental conditions as pure as possible, then the difference between the scores of the two groups should constitute a measure of the effectiveness of the experimental variable.

However, the measuring process itself may affect the characteristic being measured in any type of research which uses human subjects. And the "before-after" design is especially vulnerable to this influence. It is important to repeat that if people feel they are being used as guinea pigs, or if they feel they are being tested and must make a good impression, or if they feel "different" from other people because they are dubbed "experimental," or if they try to keep their second responses consistent with their first responses or try to vary them if they feel that variation is expected, distortion is apt to creep into the results. The "before" measuring may crystallize attitudes related to the experimental variable. The "after" measuring may yield contaminated results if the subjects reflect test fatigue or become bored or defensive. The investigator who is sensitive to these possible influences may be able to minimize their effects through careful planning. He may contrive to collect his data as part of a normal course of events and not in a special test situation. He may find ways to keep pre-measures and post-measures separate from each other and not connected in the subjects' minds. He may deliberately avoid overtesting to prevent wearing out the good will of his subjects. And he may keep situations free of threat to avoid stirring up defensive reactions.

Extensions of the control group design. To provide additional safeguards against the influence of pre-measurement, contemporaneous events, or developmental changes, more complicated extensions of the control group design have been worked out. These extensions involve adding additional control groups, each of which is handled differently. These more complicated extensions of the classical control group design are seldom used. This is quite understandable because they are obviously laborious and highly difficult to contrive. But a quick look at what we might be doing if we attempted to provide for "ideal" controls should illustrate the necessity for securing the best control group possible, even if it turns out to be something considerably less than "ideal."

The first extension involves the use of two control groups. The second control group is not pre-measured but is exposed to the independent (experimental) variable and the post-measure. The first control group is both pre-measured and post-measured but is not exposed to the experimental variable. The pre-measure for the second control group is assumed to be similar to the average pre-test score for the experimental group and the first control group. This produces a second control group, with an assumed pre-measure and a post-measure, which has been exposed to the experimental variable but which in no way can reflect the influence of the pre-measurement. If the difference between the post-measure scores of the experimental group and each of the two control groups is widely divergent, it may be assumed that there is an interaction between the pre-measure and the experimental variable. Such interaction may be found by comparing the change in scores of the three groups. Of course, the three groups should be selected by random assignment or by some matching procedure to insure that they differ only by chance at the beginning of the experiment.

The second extension involves the use of three control groups. Here again, all four groups (one experimental and three control groups) should be selected in such a way that they differ only by chance in the initial stages of the experiment. The experimental and the first control group are pre-measured. The second and third control groups are not pre-measured, and their pre-test score is assumed to be similar to the average pre-test score of the experimental group and the first control group combined. The experimental variable is introduced in the experimental group and in the second control group. It is not introduced in the first and third control groups. All four groups are subjected to a post-measurement.

Change in the first control group reflects the effects of pre-measurement, contemporaneous events, and developmental changes. Change in the second control group reflects the effects of the experimental variable and of contemporaneous events and developmental changes. Change in the third control group can be subtracted from change in the second control group to gain an approximation of the effect of the experimental variable alone. This can be done because control groups two and three are similarly influenced (presence of contemporaneous events and maturational processes) with the exception of the experimental variable introduced to group two. Thus, the extent to which the change in the experimental group reflects the influence of pre-measurement, contemporaneous events, developmental changes, or the experimental variable can be determined through comparison of the changes in this group with those in the other three groups.

All of the multiple control group designs really amount to conducting simultaneously the same experiment two or more times. In essence, this amounts to two or more replications of the same experiment under different conditions. If the results of these experiments are similar, we have greater assurance that the outcome is not confounded. If the results are not similar, we can probe for factors which may have influenced them. You will recall that Kirk conducted two simultaneous experiments (community and institution experimental and contrast groups) in his study. One experiment was used as a replication of the other. This check-countercheck procedure is what is attempted in the use of multiple control groups.

The use of multiple control groups, and sometimes of even one control group, confronts the investigator with truly momentous difficulties. But believable *experimental* research has not yet discovered ways to dispense with the control group approach, difficult as it may be. Instead, efforts have been directed toward inventing designs which resemble and may substitute for the sometimes "impossible" classical experimental group-control group design. The next four designs may be viewed as examples of the classical design, modified in ways to surmount anticipated obstacles and to make otherwise "mute" question-asking palpable.

DESIGN 2: "AFTER-ONLY" STUDY WITH CONTROL GROUP

Although Design 1 is the most frequently used and approved type of design for experimental research, the post-test only design not only avoids the possible contamination of pre-test influence and permits research investigation in instances where pre-testing is not possible, but also provides for all of the essential characteristics of true experimental designs.

While the experimental and control groups are not pre-measured, randomization of assignment to the two groups is employed to assure lack of initial biases between the groups. Properly handled, randomization can suffice without pre-test to lend credence to the assumption that the experimental and control groups are "equal" before the experimental treatment. The experimental or independent variable (X) is introduced in the experimental group only, and not in the control group. Since the experimental and control groups are both randomized instead of matched, subject to uncontrollable factors and vulnerable to unplanned, contemporaneous events, the differences between the post-measure scores of the two groups should approximate a direct indication of the experimental variable's influence alone. The post-test scores of the two groups (Y_2 and Y'_2) may be compared to analyze the influence of the experimental variable (X).

In addition to satisfying the basic requirements of experimental design, the post-test only design makes it possible to conduct research projects in those circumstances where pre-testing is inadvisable or impossible. Research studies which employ as subjects very young children taken from pre-school or the primary grades frequently cannot use pre-tests which require facility with reading and writing. Also, the ages of children place rather severe limits on the length and complexity of tests which require even a minimum of sustained effort to employ language skills. And research studies which employ as subjects older and more sophisticated children, or even adults, are at a loss to pre-test when completely new and unfamiliar concepts or abilities are the subject of investigation. The numerous current studies being addressed to research on the newer media of instruction such as educational television, language laboratories, programed instruction, and the like really have no satisfactory pre-tests available. Then we frequently are faced with variables about which we are not free to make experimental changes. Relevant variables such as the father's occupation, parents' education, and the family's previous places of residence cannot very well be experimentally manipulated or conditioned by a researcher. Sometimes the most important conditioning experiences and measurements have already taken place and further pre-measures might be considered indelicate or unnecessary.

In the case of the "after-only" with control group design, we really know very little about the "before" phase. But it is possible to reconstruct it historically or to conceptualize it on the basis of available evidence. For instance, we may use existing IQ scores, school marks, age, sex, or any one of numerous existing measures, and this generally is done to provide more complete evidence. However, the basic comparison is a statistical one wherein we use various correlational and significance of differences techniques to eliminate or confirm one factor after another. Or we may choose to control the experiment statistically through the use of analysis of variance or analysis of covariance techniques.

Example of an "after-only" with control group design. An illustration of this type of design may be found in the June-July 1962 issue of *The Journal of Educational Research* in which John D. Krumboltz and Barbara Bonawitz report on "The Effect of Receiving the Confirming Response in Context in Programmed Material."[8] This study obviously was a case within which no pre-test measures would have been possible.

The experiment was conducted comparing two approaches of presenting the confirming response in a programed textbook designed to teach prospective teachers how to write valid classroom achievement tests. The "isolation" approach consisted of presenting the desired response to the stimulus frame as a single word or phrase in the traditional programed manner. The "context" approach consisted of presenting the confirming response as a complete thought, usually by inserting the desired response in a repetition of the relevant part of the stimulus frame. A control group received a completely different program.

Thirty-two subjects were randomly assigned to the three treatment groups. The subjects for the experiment were undergraduates in a course in introductory educational psychology at Michigan State University. Subjects were assigned at random to three treatment groups: the "context" group, the "isolation" group, and a control group which received a completely different program on interpreting test results. Randomization was accomplished by arranging the three types of programs in a random sequence and distributing them to the students in the order in which they were seated in the classroom.

It was evident that the basic program on writing valid test items had a desirable impact on those taking it, regardless of the method of receiving the confirming answer. The control group scored far lower on the total criterion test than did either of the other groups. A finding of much interest was that the confirming answer "in context" seems to have a desirable effect on the ability to respond correctly to application questions.

Advantages and limitations. Although this design shares a problem encountered in all research with people (that the measurement procedures themselves may alter the characteristics they are supposed to measure) the problem is vastly less serious here than in "before-after" studies. The assumption is made that both groups are subject to similar maturational processes and to the same uncontrollable external events between the time of selection and the time at which the measure on the dependent variable (Y_2) is taken. If this assumption is justified, the position of the two groups on Y_2 at the close of the experiment includes events and processes that have affected both groups. The difference between Y_2 and Y'_2 may be taken as an indication of the experimental treatment.

Another advantage of this design for studies conducted in schools has already been noted but should be stated again because of its importance. A large number of our problems and questions related to schools have antecedents which are beyond our control and which already exist. Education is bound to family and community factors which come with the child to his school and to his learning activities. Principals and teachers can do little, if anything, to alter the father's employment status, income, or educational level, or the mother's employment outside of the home.

Thus these and other available measures may be used in lieu of pre-tests, and the effect of the experimental variable, which we *can* manipulate, may still be investigated.

The major limitation of this design is, of course, the problem of the missing "before's." Design 2 fills a need for handling the difficult or impossible "before" measuring, and in addition is appropriate in all cases where Design 1 might be used.

DESIGN 3: "BEFORE-AFTER" STUDY WITH A SINGLE GROUP

In many practical situations, finding a control group may prove to be not only difficult but also administratively impossible. This third design gets around the problem quite neatly by having each subject serve as his own control. The difference between the pre-measure on the dependent variable (Y_1) and the post-measure on the dependent variable (Y_2) is taken as a measure of the effect of the independent variable (X).

The single group "before-after" experiment is sometimes called the "successional" experiment and is often used in laboratory research. No technique is more common in the total array of research procedures than the before-and-after measurement of a variable to test the effect of a stimulus, an event, or a change which has been introduced between the first and second measurements. In the instructional setting, the range of change producing factors which are the subject of study is very broad, indeed. Administrators and teachers can easily and naturally manipulate these factors as experimental variables, as in the study of the effect of special concrete materials on learning and retention in arithmetic. At times, to test the lasting effect of a change, more than two measurements are taken. The measurements are spaced over a continuing period rather than on either side of an experimental variable, and the intent becomes truly "successional." Other variations of this sort of running experiment, of course, are available.

Example of a "before-after" study with a single group. Perhaps the best known example of a single group "before-after" experiment may be found in the book *Management and the Worker.*[9] This was the famous study carried out by Mayo, Roethlisberger, Dickson, and others at the Hawthorne plant of Western Electric Company in Chicago. It was one of several so-called illumination studies common to the social research of the thirties.

The investigators had initially come to the prediction that the introduction of improvements in the physical conditions of work would result in increased productivity of an experimental group engaged in the production of telephone relay assemblies. The "before" period consisted of intensive observation in which production and behavioral norms were identified—variation in output, interruptions, communication patterns, production rate, percentage of telephone relays rejected because of poor quality, and so forth. At subsequent intervals, the variables to be measured were introduced. The "after" phase consisted of measuring the differences in production and in the behavioral norms which, interestingly enough, impressively followed the original prediction.

But the most interesting finding was what has come to be known among researchers in the behavioral sciences as the "Hawthorne effect." In order to manipulate more exactly the physical factors affecting production, the experimenters had set up a special experimental room for the girls who were wiring the telephone relays. This wiring room was separated from the rest of the factory and, as a result, the workers became the center of attention. Furthermore, they received special attention from the outside experimenters and from the management of the plant. Careful studies of this wiring group showed marked increases in production which were related only to the special social position and social treatment they received. The fact that these workers were made the center of attention thwarted the intent of the researchers to "control" the important variables, and almost any sort of change introduced into the conditions of work resulted in increased output and improved behavior.

The major trouble with the Hawthorne study was the absence of a control group for comparison. It was possible to assume that before the relevant variables were introduced, the experimental group was the same as the control. But there was no way to tell what might have happened to the control group if there had been *no* introduction of the selected variables. Of course, the experimental group in the Hawthorne study was placed in a special position which differed from customary work conditions, and this probably set the subjects apart in their own minds. School children, accustomed to the introduction of differing instructional procedures and the subsequent testing of results, may be somewhat immune to the danger of being made special, provided that the general conditions of their schoolwork are not suddenly and drastically altered. Nevertheless, although we may sometimes be unable to contrive a separate control group and "measure the effect of the non-stimulus," we must at least be aware of the dangers inherent in this gap.

Advantages and limitations. One of the major advantages of this design is its sheer convenience and feasibility. If research studies of local school problems depended ultimately on separate control and experimental groups, vastly fewer studies would ever be attempted. And at the very least, it is better for questions to be asked and answers to be sought systematically than for nothing to be attempted because of formidable impedimenta such as finding "matched" control groups or locating a sufficient number of subjects to make possible a random assignment to control and experimental groups. The ordinary classroom group is composed of twenty to thirty students, and this may be the total number of subjects a teacher can have at his disposal. The invention of a design pattern where each subject acts as his own control in effect doubles the number of subjects available, and no more precise matching is possible.

The use of this design may be justified when the investigator has good reason to believe 1) that the "before" measure will not in some way affect the response to the experimental treatment, and 2) that there are not likely to be any other influences, besides the experimental treatment, during the course of the experiment which might affect the subjects' responses at the time of the "after" measurement. In order to be reasonably sure that such assumptions are justified, the investigator must have considerable knowledge of the probable effects of his measurements and of conditions other than the experimental treatment that are likely to influence the dependent variable. This is fortunately true in the case of individual school environments where principals and teachers are apt to approach problems with an already impressive array of knowledge about the subjects and where they normally exercise considerable control over the planning and conduct of the learning environment.

Another advantage can be found in small *N* studies (25 or so subjects). In contrast to large-scale experiments, the smaller number of subjects makes possible deeper and more complete study of individual subjects. In addition to analyses of a group of subjects—for example, statistical analyses—individual case studies can be completed. Kirk did this effectively in his study. In truth, the staff members of individual schools are in a unique position to exercise a kind of control and completeness that is simply not available to outside researchers.

Sometimes, the "before-after" design used with a single group of subjects is criticized because it is alleged that the findings cannot be generalized to other groups. In the case of the schools, this may not strictly be the case. The teacher-researcher wants to engage in investigations that will help him make better decisions and improve his practices *now*. He also hopes that his findings will apply to the children he will teach in the future—probably in the same school and at the same grade level. Whether they will apply, of course, depends upon the extent to which his students *now* represent a random sample of the population he will have in the future.

Stephen M. Corey, in his book *Action Research to Improve School Practices,* presents some interesting material on this point.[10] He reports on a series of statistical tests of the null hypothesis that the distribution of scores on individual classes and the distribution for the total population of *N* classes taught by the same teacher will not differ more than might be expected as a consequence of chance. Students in twelve different class sequences extending over three years, taught by twelve teachers in New York City and West Orange, New Jersey, were the subjects of the analysis. The variables studied were IQ, subject matter achievement test scores, and silent reading test scores. The chi square test of the significance of differences was applied to the distribution of scores. The probabilities resulting from thirty-eight chi square tests, with one possible exception, indicated that the students in a teacher's class at any one time could be considered to be a random sample of the total population of several consecutive classes. Of course, any teacher, if he has available the necessary scores for several consecutive classes, can apply the chi square test to determine if the differences between the classes are greater than would be expected by chance alone.

It is quite pertinent to inquire into how much confidence can be placed in the generalizations derived from a single study, even if the study's sample was randomly selected from an indefinitely large population. For example, some investigator might use a random sample of all fifth graders in the State of New York or the State of California. But teachers anywhere would do well to question whether the immediate group with which they are concerned is at all similar to the one on which the "randomly sampled" study was based. Research studies in education have long been criticized for the tendency of researchers to generalize on the slim basis of a single study completed once and never again. There is a strange lack of replication of studies in education, to the detriment of believable scientific knowledge. In this state of affairs, it might be wise for any practitioner in education to form the habit of performing check studies in his own situation before accepting generalizations derived from unreplicated studies done elsewhere.

Perhaps the major limitations of this design have been adequately noted in relation to the Hawthorne effect. The chief weakness is, of course, that there is no control group to provide information on the effect of the nonstimulus.

DESIGN 4: "AFTER-ONLY" STUDY WITHOUT CONTROLS

Perhaps a design equally common with the one just described is the single cell. This generally represents the over-all design used in school surveys, many research studies, assessment studies, status studies, and case studies. Treatment of the data is usually descriptive in nature, with only incidental attempts to attribute findings to associated variables. As such, this is actually a report on what exists at the time of the study. It is not a design used in experimental research, though it may be a necessary step in charting the territory for later experimentation.

When we have gathered adequate data about the group being studied, we can relate various characteristics and gain information about factors which may have been previously obscure. The more we study such a single cell, the more confident we become that certain forces may be influential in shaping the results we uncover. Our logical analyses may jibe with good common sense and suggest a flood of hypotheses to be tested later. Studies conducted within this design can be very useful, but they do not test reasoned solutions to conceptualized problems. Ordinarily, surveys and status studies are insufficient for experimental purposes. The chain of events associated with experimental research calls for some sort of both control and experimental groups to see 1) how the experimental variable has affected the experimental group, and 2) how the lack of, or difference in, the experimental variable has affected the control group.

Example of an "after-only" study without controls. One massive "after-only" study was conceived and conducted by the twenty-seven members of an ASCD Yearbook Committee and published as the 1958 ASCD yearbook, *A Look at Continuity in the School Program.*[11] The committee members chose to study the present situation with respect to articulation and learning continuity as viewed by children from kindergarten through the senior high school. To do this, the committee members worked out a simple interview guide to give a structure to children's reports in a group interview situation. Committee members or their representatives talked with groups of children about experiences which had helped or hindered them in their progress through school. Then they asked the individual children in each group to answer four simple questions written on a mimeographed guide sheet:

"1. Tell about anything that has happened to you which has helped you to feel that your progress was smooth and that the school helped you move along without unnecessary difficulty.

"2. Tell about anything that has happened to you which has made it difficult for you to move along smoothly through school.

"3. Tell about any experience that has been very pleasant or very unpleasant and which grew out of this question of your progress through school. It may have happened quite a while ago or lately. Write as much as you can remember about it; tell how it made you feel at the time it happened and how you feel about it now.

"4. If you have moved from one school to another, tell how you felt about moving before and after you moved. Do you feel the same way now?"[12]

The committee's plan for gathering data purposely allowed for variation in responses. Each committee member was left free to modify the interview guide form in any way that seemed desirable to him, as long as he got children to give their own free reports to the four basic questions. Particularly with young children, the interviewers had to modify and reword the questions. In some cases, it was necessary to take notes, use recording devices, or arrange for stenographic reports.

Sets of student reports were submitted by seventeen committee members from eleven states. They had collected a total of 2731 usable reports. A group of analysts (the committee chairman and twelve advanced students of the University of Alabama) reviewed the completed reports and worked out data sheets to which the material could be transferred. This procedure resulted in organizing the data under common categories regardless of where in a child's report the pertinent material was recorded. Any type of situation which represented at least one per cent of the total situations was included as a category in this grouping. The data sheets were also classified according to grade level: primary, intermediate, junior high school, senior high school, and a mixed group for which the grade level at the time of the incident was not indicated. A total of 4197 data sheets were used in the final summarization.

The completed study organized the data by grade level groups around fourteen categories of situations. Eleven tables were required to present the summarized data. The data in the tables were grouped according to frequencies, percentages, and rank order of positive and negative reactions.

The fourteen categories of responses to the initial four guide sheet questions turned out to be: 1) moving to new school community, 2) teacher behavior, 3) subject matter, 4) moving to next school level, 5) smooth progress, 6) extracurricular activities, 7) differences in teaching methods, 8) illness, 9) rewards, 10) punishments, 11) promotion, 12) grading, 13) retention, and 14) accidents. Interestingly enough, the percentage of negative responses reflected on the data sheets was greater than the percentage of positive responses.

The writers of this yearbook were explicit and positive in their insistence that the study was merely an introductory and exploratory treatment of a subject which is in need of much study, research, and evaluation.

Advantages and limitations. The survey or descriptive study is a process for learning pertinent information about an existing situation. The principal devices for gathering data from people involved are the interview, the questionnaire, and summaries of existing documents which count and enumerate phenomena, such as those which are published by the Bureau of the Census and by the Research Division of the National Education Association.

The survey frequently becomes more than a mere fact-finding device. It may result in important hypotheses or conclusions that help to solve current problems, and it may provide basic information for comparison studies and for identifying trends. Surveys completed five or ten years ago may be repeated, and something will be learned about changes which have taken place in the meantime. School district surveys have long served certain needs in education by providing previously lacking descriptions of curricula, attendance centers, citizen opinion, and physical facilities. They have also helped to pool divergent ideas, techniques, and bits of information, thus throwing light upon existing conditions in need of change and improvement.

Most times, it is far easier to obtain data through single cell studies than it is to draw valid conclusions from the facts discovered. Studies of this type tend to have a larger scope and to range more freely than experimental studies with control groups. Surveys frequently tend to be composed of a loose confederation of several single cells related or unrelated to each other. These characteristics lend to the studies the impressions of size and convincingness. And herein lies perhaps the major limitation of single cell studies. The freedom to range more widely and not to be strictly bound by the requirements of experimental research also makes impossible definite problem solution, prediction based on theory, and probability inference. In spite of feelings of certainty which may be drawn from single cell studies, it is important to note that such feelings or convictions are not demonstrations. The very real and necessary contributions to be made through pre-research studies, case studies, surveys, status studies, and other forms of enumerative investigations should not be diluted by reading into them more than can be demonstrated.

DESIGN 5: EX POST FACTO EXPERIMENT

The basic pattern of the classical experimental design is one of comparing the result of the experimental variable upon an experimental group with the result of *not* introducing that variable into a comparable group under the same conditions. Most frequently, this involves a prediction of what will be the difference in results between the two groups, i.e., it is used as a projective design which looks forward in time. However, the classical logical pattern need not be confined to projective designs. It can also be applied to problems which look from the past to the present rather than being oriented toward the future. This is known as ex post facto research.

The major characteristic of this procedure is that the investigator can control the crucial variables only by selecting those which have already been recorded. The investigator may locate a group of subjects which has been exposed to the type of experience in which he is interested and another group, similar in other relevant respects, which has not been exposed to such an experience. He then compares the groups on the basis of present performance. In basic logic, however, projective and ex post facto experiments are the same. The purpose of both is to compare two groups similar on all relevant characteristics save one in order to measure the effect of that characteristic.

In the ex post facto experiment, we manipulate pieces of paper instead of experimental conditions. That is, we manipulate existing records which symbolize the behavior upon which the experiment is focused. Control is achieved by matching records in such a way that the control and experimental groups are similar in relation to all but the crucial variable. The records are then followed through by measuring the consequences as observed in present behavior.

The teacher-investigator may not be in a position to test a hypothesis by assigning subjects to different conditions in which he directly controls the experimental variable. For instance, there very well may be a limit to which certain types of instruction can be justifiably withheld from a group of control students. Or the teacher may be interested in exploring the effect of some long-range school practice but simply does not have time to wait several years for maturational effects to become evident. Consider, for example, the research that has been done on incidental (activity type) instruction in arithmetic. The extent of incidental instruction compared with formal instruction has not been controlled, as it would be in a "purely" experimental study, by assigning different individuals to classroom groups where instruction proceeds according to different basic curriculum patterns. Rather, records have been secured which symbolize the arithmetical behavior of students in activity (experimental) schools and students in formal (control) schools. The relation between type of instruction and performance in arithmetic is then computed. And for many reasons, it is not likely that the basic design of a school curriculum will be manipulated to suit experimental conditions.

The ex post facto pattern involves the same problems of matching that are present in projective designs. However, it has two additional problems that are uniquely inherent: 1) It is necessary to find adequate records after a lapse of some years; and 2) Problems are restricted to those areas for which adequate records exist and are available to the investigator. If important material was not recorded, the needed information cannot be resurrected just for the purpose of study.

However, teachers find themselves in an unusually advantageous position to make ready use of complete records that go back in time. Administratively, schools have become avid storers of information: academic marks, intelligence quotients, health charts, attendance records, height and weight charts, aptitude scores, citizenship marks, parental information, judgmental descriptions by teachers, personality scores, achievement scores, sociograms, and a host of additional measurements. In fact, there are very few other sources of meticulously kept records which cover so wide a range of the behaviors of people over sustained periods of years. Theoretically, it is possible to reconstruct Johnny at any particular time—in terms of his father's occupation, the story of his physical health, his rate of physical growth, his intellectual achievements, his sequential progress through school, what his teachers thought of him, how he got along with playmates, what his talents were, the developing pattern of his abilities, and many more qualities belonging to Johnny and classroom groups around him at various times.

Frequently, the investigator may not find it possible to assign individuals to different groups, one of which will have the experimental treatment and one of which will not. An alternative solution is to find two comparable groups which will be, or have been, exposed to experiences which differ in relation to the crucial variable of interest. This alternative is a very common procedure in most types of research studies and serves very well in the case of ex post facto experiments.

Example of an ex post facto experiment. A particularly ingenious scheme was used in one study, "The Grouped and the Ungrouped," by Jo Taylor Auld.[13] This study set out to investigate the effects of traditional "high," "average," and "low" ability grouping in elementary schools. It could not have been conducted at all if there had not been already existing records which could be manipulated on an ex post facto basis.

Two elementary schools in a South Carolina city were included in this study. The neighborhoods served by the two schools were alike in terms of housing, community facilities, churches, etc. In each school, the principal had held the post for more than fifteen years. The number, age, and experience of faculty members were approximately the same. The facilities in the schools were equal, as were the organized play programs, library periods, and so forth. Being in the same city system, the schools operated under the same policies and followed the same testing procedures.

In School A (the experimental school) there were thirty-one boys and twenty-seven girls in the sixth grade who had been in attendance there since they entered the first grade. These children had been grouped in "high," "average," and "low" groups when they entered the second grade. The grouping had been based on teacher judgment and on reading ability as measured by the California Achievement Test.

In School B (the control school) there were twenty-nine boys and twenty-seven girls enrolled in the sixth grade who had been attending the school since they entered the first grade. These children had not been grouped in any manner throughout their six years in school. In other words, they had been in heterogeneous classes throughout their school experience.

In each school, there was an active PTA with a membership of 100 per cent of the families. The families were much alike in the two schools with regard to their socio-economic status. More fathers in School B were engaged in professional occupations, and slightly more mothers in School A were employed in positions outside the home.

The median intelligence quotient of the group of children in School A was 107, with a range from 87 to 134. In School B, the median was 110, with a range from 87 to 133.

During the 1959-60 session, Auld studied the test results obtained when the children in School B were in the first grade. On the basis of the reading test and the recorded teacher evaluations, she arranged their names in three groups—"high," "average," and "low"—in the same manner that the children would have been grouped for instructional purposes if they had been enrolled in School A when they entered second grade. Thus, she was able to establish comparable groups in the two schools and study the differences, if any, that existed in the students after more than four years of schooling under different systems of grouping.

The following findings are based upon the results of the Metropolitan

Achievement Test (Intermediate Battery VI) which was administered to the children of both schools.

1. There were no statistically significant differences in the performance of the students in Schools A and B who were in the "high" or "average" groups.

2. Among the "low" group, those in School B demonstrated significantly more achievement than those in School A. This difference was quite marked, with the School B students earning a mean grade equivalent of 6.3 whereas those in School A earned a mean of only 5.0. This difference was significant at the .01 level of confidence.

3. Achievement in reading, arithmetic, and language on the part of "average" and "low" groups was higher in School B where the students had not been grouped for instructional purposes. The data indicate that at least among these students in these two schools, grouping over approximately a four-year period apparently hindered the achievement of the average and below average pupils, and it did not provide advantages for the above average or "high" pupils.

Advantages and limitations. The major advantages of the ex post facto experiment cluster around the simple fact that it makes it possible to seek answers to certain kinds of questions that otherwise would go unanswered. The ex post facto pattern in effect compresses time, permitting study of the effect of many years' experience now, rather than waiting for the experience to happen. In a very real sense, this expresses a common and regular responsibility of all schools everywhere. Measurements are made and records are kept to predict the probable success of students in college, in jobs, and as citizens. Grade point averages, academic experiences, and citizenship marks are used as predictors of future abilities and accomplishments. These records, if carefully secured and preserved, can provide the needed data to check the educational procedures which gave rise to them in the first place.

Of course, the investigator will find it necessary to account for several factors which might operate to confound the data. He will want to be sure that the individuals who have undergone different experiences were comparable before they were exposed to the different experiences in question. He will want to be sure that the time order and intensity of exposure to contrasting variables are similar. He will account for some of the extraneous variables which might have operated during the course of the experiment. And he will avoid such obviously confounding factors as comparing an assigned group with a self-selected group. The Auld study, for example, accounted for factors such as these. But it must be remembered that her study could not do more than approximate the control group-experimental group design.

The major limitation of the ex post facto experiment is that it cannot provide safeguards as adequate as those given by random assignment of individuals to experimental and control groups, control over the extraneous variables that might operate during the course of the experiment, and direct manipulation of the experimental variable. In addition, the experimenter is limited by the degree of thoroughness and accuracy which unknown and past record keepers have displayed. And of real concern and frustration is the danger of encountering "blank spots" where some crucial data may not have been recorded at all or may have been recorded in an unusable manner.

In the same sense that the creative investigator may view himself as one who observes what happens when he treats the world as if it would operate as he hypothesizes, so he may view himself as one who can plan (design) events to facilitate the operation of his hypotheses. He devises ways to control and manipulate conditions to make observations of his hypotheses-in-action clearer and sharper. This calls for ingenuity and invention. It also calls for the recognition of opportunities inherent in the original problem and its hypotheses in order to exploit new ways to plan for collecting data or to modify already known designs which have been successfully used elsewhere. For instance, Auld employed unusual insight in employing the ex post facto records in one heterogeneously grouped school to arrange the children's names in the same manner that they would have been arranged had they been in a homogeneously grouped school. In this manner, she contrived a way to control and manipulate conditions so that her hypothesis could be tested.

It would be a mistake to regard the various designs for conducting tests of hypotheses as complete, final, or restrictive. The world is wide open for new and thoughtful ways to plan the conduct of problem solving. And theoretically, at least, there should be as many different ways to plan as there are different problems to be solved and different people to solve them.

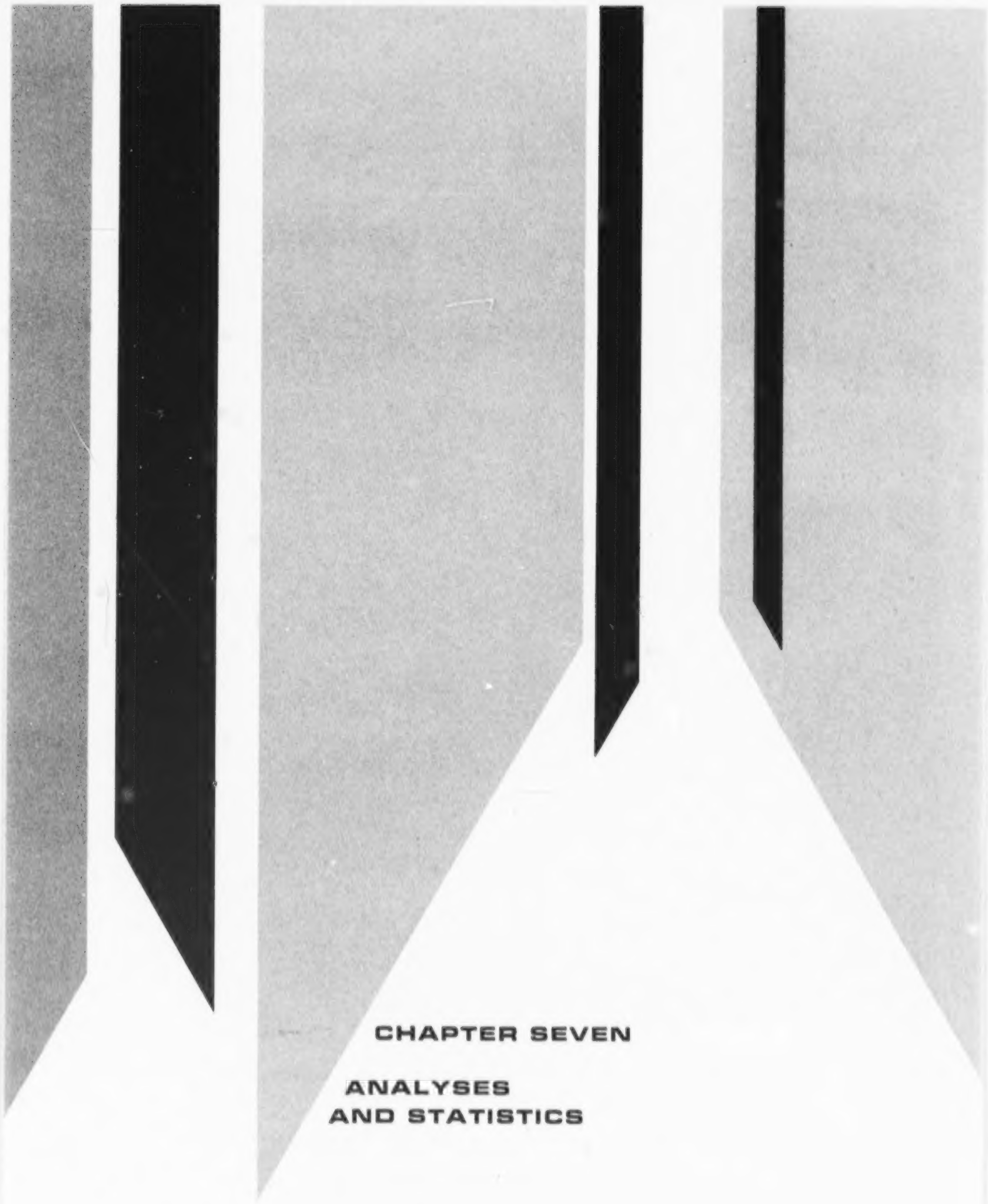
This chapter has presented only five types of "standard" design used in research. Many more are described in the literature on research methods. Some are highly complicated and include unduly difficult or impossible requirements for people engaged in actual situations. Others are purely statistical, focusing chiefly on mathematical models. But the problem of design is mainly logical in nature and is only secondarily statistical for the ordinary investigator. School people would do well to expand gradu-

Five Types of Design	"Before" measurement	Exposure to experimental variable	Exposure to uncontrolled events	"After" measurement	Analysis of data
I.					
"Before-After" with control group					
Exp. Grp.	Yes (Y_1)	Yes	Yes	Yes (Y_2)	$d = Y_2 - Y_1$
Con. Grp.	Yes (Y'_1)	No	Yes	Yes (Y'_2)	$d' = Y'_2 - Y'_1$
II.					
"After-Only" with control group					
Exp. Grp.	No	Yes	Yes	Yes (Y_2)	
Con. Grp.	No	No	Yes	Yes (Y'_2)	$d = Y_2 - Y'_2$
III.					
"Before-After" with single group					
Exp. Grp.	Yes (Y_1)	Yes	Yes	Yes (Y_2)	$d = Y_2 - Y_1$
IV.					
"After-Only" without controls					
Exp. Grp.	No	Yes	Yes	Yes (Y_2)	descriptive Y_2
V.					
Ex Post Facto Experiment					
Exp. Grp.	Yes (Y_1)	Yes	Yes	Yes (Y_2)	$d = Y_2 - Y_1$
Con. Grp.	Yes (Y'_1)	No	Yes	Yes (Y'_2)	$d = Y'_2 - Y'_1$

A variable is a quantity whose value may change. When a change in one quantity is caused by a change in another, the value of the first depends on the value of the second; so the first is called the *dependent* variable (Y), and the second the *independent* variable (X).

Figure 5. Types of Experimental Designs

ally their knowledge of the alternatives suggested by any available design pattern and then choose the ones most feasible, promising, and flexible for their own research interests.



CHAPTER SEVEN

ANALYSES AND STATISTICS

The data resulting from any experiment are usually a collection of observations or measurements. We have seen that these data are more likely to yield credible information and apply to the answers that we seek when the experiment is guided by adroit hypothesizing, careful sampling, and insightful pre-planning or designing. In addition, when we considered variables and probability, we saw that the testing of hypotheses requires four types of evidence: 1) concomitant variation, or evidence that the independent variable (cause) and the dependent variable (effect) are associated; 2) time sequence, or evidence that the dependent variable did not occur before the independent variable; 3) internal control, or evidence ruling out other factors as possible influences; and 4) probability, or allowance for the operations of chance or sampling error—instead of the deliberately introduced independent variable—to have produced the effects that we observe in an actual experiment.

During and after the collection of data, it is necessary to analyze them in terms of the four types of evidence which lend reliability to the conclusions reached. Logical analyses apply quite handily to the first three types of evidence. But the fourth type of evidence, allowing for the operations of chance, can very seldom be ascertained by simple direct inspection of the data or by logical analyses of them. Statistics provides the methodology whereby this can be done.

NONPARAMETRIC AND PARAMETRIC METHODS

The fundamental statistical problem arises when we try to see if one variable produces a change in another variable in a situation where the latter *spontaneously* changes due to the influence of other causal factors which are commonly termed "chance." In other words, starting with a mathematical model for random variation, we ask what the chances are of drawing two separate samples whose mean or median difference is as great or greater than the difference obtained from an actual experiment. If the chances are small (say one in a hundred) of obtaining a difference of this magnitude from two random samples drawn from the same parent population, we may tend to regard favorably the likelihood that the actual difference is due to the effect of the independent variable which was deliberately introduced into the experiment. But if this likelihood is considerably greater (say fifteen, twenty, or more in a hundred), we dismiss the difference as easily due to chance. Of course, this reasoning involves the assumption that the variance in the population or samples with which we are dealing in an actual experiment is normally distributed, as it is in the mathematical model. If it is not—and it often is not—then the justification for the comparison is seriously questionable.

Because such assumptions are often unsuitable in practice, and because the requisite conditions on the assumption of "normalcy" in the distribution of sample or population scores either are not met or are untestable, increasing attention has recently been paid to the newer statistical tests that do not depend upon such assumptions and conditions. These are the *nonparametric* or "distribution-free" tests. Some nonparametric techniques are called "ranking tests" or "order tests," and their titles suggest their characteristics.

The customary "standard" techniques of statistical inference are based on several strong assumptions about the nature of the population from which the scores were drawn. For example, a technique of inference may be based on the assumption (seldom tested) that the scores were drawn from a normally distributed (bell-shaped) population. These population values are called "parameters"—a mathematical term indicating variables entering into any distribution where the possible values of the variable correspond to the distribution. Because of this, the usual statistical techniques are called *parametric*. That is, parametric techniques are tied closely to the shape of the distribution of scores or observations. When using parametric tests, we arrive at conclusions which may be valid *if the assumptions regarding the shape of the population are valid.*

Nonparametric or distribution-free techniques require fewer qualifications and assumptions regarding the shape of the population. In testing hypotheses about a control group and an experimental group, the comparison is between distributions and not between parameters. When using nonparametric tests, we arrive at conclusions which may be valid *regardless of the shape of the population values.*

Several disadvantages and advantages of nonparametric methods have been pointed to by Lincoln E. Moses.[14] Paraphrasing the Moses treatment, we may compare on several counts distribution-free techniques and normal-distribution techniques. This may provide a rather quick summary of reasons why nonparametric techniques seem reasonably well fitted to research projects undertaken in individual school systems, individual schools, and individual classrooms.

Disadvantages of nonparametric methods. 1. Nonparametric tests applied to normal data—rather than normal-theory tests—are wasteful of data. The degree of wastefulness is measured by the "power efficiency" of a nonparametric test when compared with a similar parametric test. "Power" can be defined as the probability of rejecting the null hypothesis when the alternative (research) hypothesis is true.

If, for example, a test has 80 per cent efficiency, this means that *where the data are from a normal distribution*, the appropriate classical test would be just as effective with a sample of 20 per cent smaller size. In this way, "efficiency" (a mathematical test of statistical tests, both parametric and nonparametric) expresses the relative merits of the nonparametric test and the classical test under conditions *where the normal test is correct*. It does not, however, tell us how the tests will compare on non-normal data.

2. For large samples, some of the nonparametric methods require a great amount of labor.

3. Writing in 1952 and 1954, Moses listed as another disadvantage the fact that the nonparametric tests and tables of significance values were widely scattered in the periodical literature. Since 1954 this objection has been largely overcome through the publication of several quite usable books which contain both the methodology and the necessary tables. Among these are:

Wilfred J. Dixon and Frank J. Massey, Jr. *Introduction to Statistical Analysis.* New York: McGraw-Hill, 1957.

Norville M. Downie and Robert W. Heath. *Basic Statistical Methods.* New York: Harper & Brothers, 1959.

George A. Ferguson. *Statistical Analysis in Psychology and Education.* New York: McGraw-Hill, 1959.

Sidney Siegel. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill, 1956. (This book, entirely given over to nonparametric tests, contains a very large collection of appropriate tables of significance values.)

Merle W. Tate and Richard Clelland. *Nonparametric and Shortcut Statistics.* Danville, Illinois: Interstate Printers and Publishers, 1957.

The appearance of the literature on nonparametric statistics has been matched by the spread of knowledge and the use of distribution-free methods. Just a decade or so ago, nonparametric techniques seemed to be the exclusive property of mathematical statisticians in both this country and England. Since that time, researchers in many theoretical and applied fields of knowledge—psychology, agriculture, sociology, biology, education—have learned to use these newer techniques.

Advantages of nonparametric methods. 1. If samples are very small (as small as six or fewer members), there is no alternative to a nonparametric test.

2. If the sample consists of observations from several different populations (and perhaps a different number of cases in each subsample), there may be a suitable nonparametric treatment.

3. In certain cases, data can only be taken as "better" or "worse," or "gain" or "loss." That is, an observation can only be expressed as a plus or minus. Obviously, the classical tests are not applicable to such data, but some distribution-free tests are.

4. If the data are inherently in the nature of ranks, not measurements, they can be treated directly by nonparametric methods without assuming some special shape for the underlying distribution.

5. Whatever may be the shape of the distribution from which the sample has been drawn, a nonparametric test of a specified significance level actually has that exact significance level.

6. Nonparametric methods are usually much more understandable and much easier to apply than the classical techniques. Pencil and scratch paper will do quite adequately for most calculations, and there is much less need for electronic computers. Possibly this is one of the strongest arguments for these methods in the case of investigators who must spend something less than full time on research activities and without the aid of expensive calculators.

MEASUREMENT SCALES

At this point, the nature of the data with which the behavioral sciences, including education, must deal is very much to the point. The operations allowable on a given set of scores are dependent on the *level* of measurement which has been achieved. The theory of measurement consists of a series of levels or *scales* of measurement: 1) nominal, 2) ordinal, 3) interval, and 4) ratio.

The *nominal scale* consists of numbers or other symbols which are used to identify and name groups to which various persons or objects belong. Identifying boys and girls as *B* and *G* (or 1 and 2) is an example of using a nominal scale.

The *ordinal scale* assigns numerals to objects which are rank-ordered with respect to some characteristic. The persons or objects can be ranked starting with the highest score, and then proceeding to the next highest, and so on until the lowest score is reached. Or the ranking can start with the lowest score, proceed to the next lowest, and so on until the highest score is reached. Thus, the rank of students in a graduating class, or the order of students from high to low on examination questions correctly answered, gives us an ordinal scale.

The *interval scale* defines a unit of measurement such that a difference of one scale value is equal to any other one scale value. Dates on a calendar, degrees on a thermometer, or carefully constructed standard achievement test scores may approximate an interval scale.

The *ratio scale* is an interval scale which starts at an absolute zero point, so that it is correct to speak of one scale value as being twice, or three times, etc., as large as another. Thus, weight of objects expressed as pounds, or ounces, or grams has a true zero point, and the ratio between any two points is independent of the unit of measurement. But IQ scores, for example, are not ratio measurements since there is no such quantity as 0 IQ, and an IQ of 120 does not simply indicate twice as much intelligence as an IQ of 60.

The major point to be served by this rapid consideration of measurement scales is that typically in the behavioral sciences, our measurements are unavoidably crude and inexact—compared with the physical sciences, for instance. It may be illustrative to consider that ratio scales are rarely achieved in the behavioral sciences. Thus, research in education is left with three types of usable scales: nominal, ordinal, and interval. Nonparametric tests are most useful with the first two types of measurements, in the order listed, and to some extent with the third type. Parametric tests are most useful with interval scale level of measurements and progressively less useful with ordinal and nominal levels of measurement. To the extent that the bulk of our judgments, scores, and measurements in education cluster around the nominal and ordinal scales, distribution-free techniques increase our research equipment and flexibility.

STATING AND EVALUATING THE NULL HYPOTHESIS

When we have secured the desired evidence, as expressed in terms of any measurement scale, we then proceed to evaluate it in regard to levels of significance. We will next consider what is meant by "significance," and how this level is determined in actual situations.

It has been stressed repeatedly in this publication that the job of statistical studies is to handle the problem of the *null hypothesis* (the hypothesis that the results obtained for the difference between two or more groups is *not* due to the deliberately introduced independent variable, but is due to an error or accident of sampling). Evidence is used not to "prove" some positive hypothesis (a logical impossibility) but to progressively "disprove" or discredit the null hypothesis.

Levels of significance. In the application of a statistical test of the evidence, we are interested in determining if the observed deviation between two groups is too large to be accounted for by chance. The statistical model that we select will enable us to calculate the probability that the results we observe might be attributed to chance. The statistical tables which accompany the model will allow us to read directly the "significance" of the evidence produced by an actual experiment. If the probability is small that chance variation might have produced the results we observe (say 5 out of 100 repeated times) then we can reject the null hypothesis (H_0) and accept the research or alternative hypothesis (H_1). On the other hand, if the probability is too large that chance might have produced our observed results (say 25 out of 100 repeated times) then we fail to reject the null hypothesis and, of course, we cannot accept the research hypothesis.

Customarily, the research worker states ahead of time at what level of significance he will reject his null hypothesis—the amount of risk he is willing to take. If he decides in advance that all decisions will be made at the 1 per cent level, he merely disregards other results, such as a difference being significant at the 5 or 10 per cent level. It should be noted that there is nothing sacred, other than custom, about the 1 and 5 per cent levels. There may be situations in which an individual is willing to operate at the 10 per cent level or more. Note, however, that when we reject the null hypothesis at the 1 per cent level, we have 1 chance in 100 of being wrong in that decision; at the 5 per cent level, we have 5 chances in 100; and at the 10 per cent level, we have 10 chances in 100 of being wrong in the decision.

Decision theory. The choosing of the level at which we will reject the null hypothesis is known as decision theory. There is a rather extensive, and sometimes puzzling, body of literature dealing with the cautions advisable when deciding that a null hypothesis is false (and consequently rejecting it) or deciding that it is true (and failing to reject it). But the operational qualities of decision making can be related closely to the consequences of accepting and acting upon the knowledge derived from research projects, and this is not obscure, nor is it mathematical or puzzling.

In this context, deciding is contingent upon the nature of the *risks*, or *values*, which are involved. For example, much interest has been evidenced in regard to "curative" drugs for cancer. It is quite understandable that people who might prescribe and use such drugs insist upon experimental evidence secured at a very high level of significance, such as .001 (1 chance in 1,000 of being wrong), or at an even more extreme level. The consequences involved in the general use of such a medicine, tested under conditions where assumed curative properties might too easily be due to chance, are that other treatments tend to be stopped, and the results may be disastrous. On the other hand, an experiment concerned with temporary or minor changes in diet can safely afford to be based on a decision to use a much larger significance level—say the 5 or 10 per cent level. The chances are reasonably good that something useful will be learned and, at the very least, serious and lasting injury will not be the result.

In other words, "truth" is a relative matter. The evidence at our disposal is always partial, incomplete, and to some extent, tentative. This is part of the reason for saying that "proof" of a positive hypothesis is a logical impossibility; there is always some room for doubt, no matter how small. Also, what is "good enough" for one purpose may not at all be sufficient for another. And different persons who intend to use the results of certain research projects may be willing to accept or not accept the findings as "good enough" at differing levels of significance. An investigator cannot foresee all of the situations and circumstances in which his findings might be relevant. Since this is true, he may be exceeding his responsibility and right to pre-judge the matter for other people by setting a decision point in advance of experimentation—a decision point on one side of which the results will be declared "significant" and on the other side, "not significant." Of course, the experimenter is free to use his findings as he sees fit, but he would show more consideration and courtesy if he would simply indicate his findings as "significant at the ......% level," and leave decisions regarding the *use* of his findings to those interested in reading his report.

Ways to handle the null hypothesis. In stating and evaluating the null hypothesis, there is one more area of information that is essential to statistical testing and to reading the tables of significance which have been prepared for the various statistical models. This area of information is concerned with two different ways to state and handle the null hypothesis, depending upon the researcher's purpose and state of knowledge on the problem being investigated. The two ways for handling the null hypothesis may be illustrated through reference to two broad sets of conditions surrounding all research investigations.

The first condition involves those instances wherein we are able to predict what kind of differences we will find between experimental and control groups—for instance, that experimental scores will be higher or lower than control scores. The research (alternative) hypothesis makes the prediction, say, that experimental group scores will be higher. The appropriate null hypothesis would be that experimental group scores will be equal to or less than those of the control group.

The second condition involves those instances in which we are

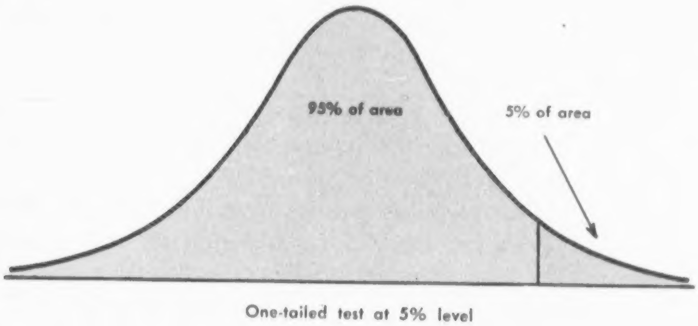

95% of area
5% of area
One-tailed test at 5% level

Figure 6. Area of Rejection for a One-tailed Test

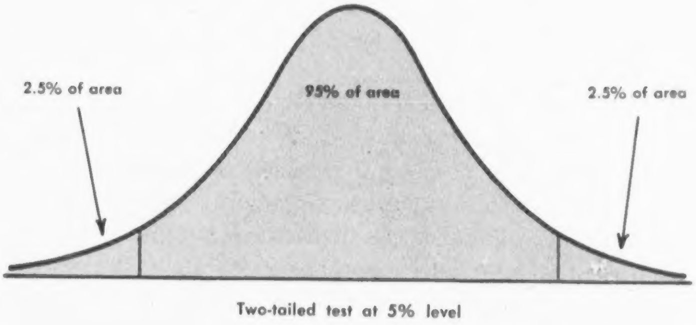

2.5% of area
95% of area
2.5% of area
Two-tailed test at 5% level

Figure 7. Area of Rejection for a Two-tailed Test

merely trying to find out if there *is* a difference between the two groups. Perhaps we do not know enough to predict the direction of the difference, or perhaps our interests will be satisfied if we simply find that the two groups differ in any respect. This time our research hypothesis would be that the experimental group scores will differ from the control group scores (either higher or lower). The appropriate null hypothesis would be that of no difference.

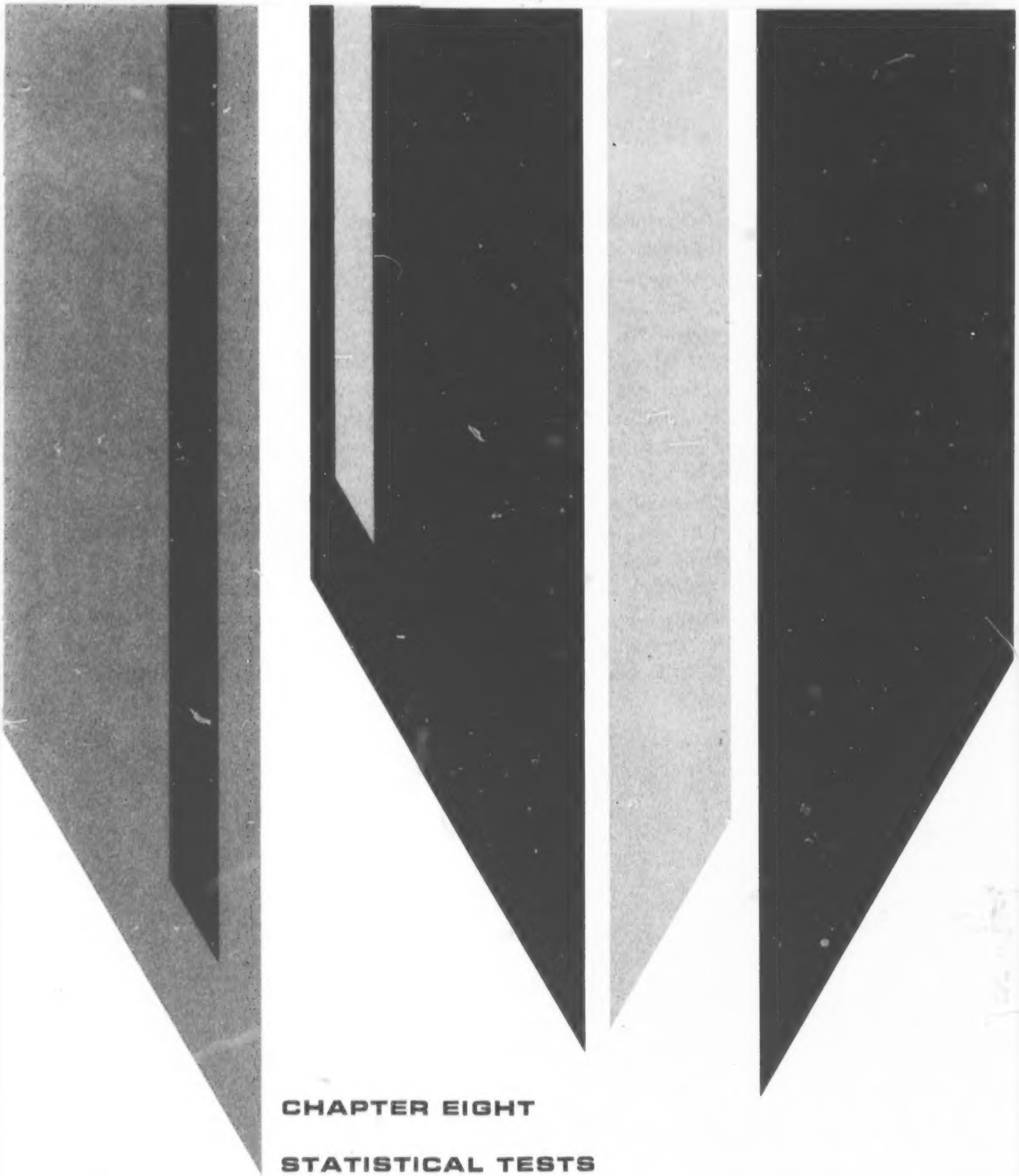
In those instances where, on the basis of theory, previous experience, or other cues, we are able to predict the direction of difference, we make what is called a *one-tailed test*. In those instances where we are not able or do not wish to predict more than that there will be a difference, we make what is called a *two-tailed test*. These two terms refer to the tails of the distribution, as shown in Figures 6 and 7.

Suppose that we set up an experiment using a new reading technique. We select two groups: the control group which is taught by one of the conventional reading methods, and the experimental group which is taught by the new method. We have reason to feel, on the basis of newly developed theory and prior research findings, that the experimental group will score higher on a post-test in reading which is given following a period of instruction. The null hypothesis that we test is that the experimental group will have scores equal to or less than the control group. But suppose we are working in an instance in which there is no relevant theory and in which there are no reports of prior research studies addressed to this particular problem. Yet we feel that the new method and the conventional method will not yield identical results. Our interest becomes one of testing for differences between two methods—in any direction. This time the null hypothesis is that there will be no difference between experimental and control group scores.

When we use a test statistic to discredit a null hypothesis, we are using a sampling distribution. This total distribution includes *all* possible values that a test statistic can take under the null hypothesis and is represented by the curve of that distribution. We are interested in determining whether the sample we actually observe will yield a value which is among the values associated with a null hypothesis, when it is true. If our observed sample value is located in a very small area of the entire space included under the curve of the sampling distribution, then we have reason for rejection. This small area is known as the *area of rejection*.

For a one-tailed test at the 5 per cent level, we have all 5 per cent of the area under the curve in one tail. In using a two-tailed test, we use both ends of the curve. At the 5 per cent level, we have 2.5 per cent of the area in each of the two tails, for a total of 5 per cent.

The tables of significance are worked out for both one-tailed and two-tailed tests. From them, it is possible to read directly the values required for significance at several levels (.01, .02, .05, .10, .20, etc.) These levels correspond to the various areas of rejection under the curve of the sampling distribution. When the investigator has decided upon his research design and upon his hypotheses to test, he is able rather quickly to determine the necessary significance values for either one-tailed or two-tailed tests of the null hypothesis.



CHAPTER EIGHT

STATISTICAL TESTS

Nonparametric tests represent direct applications of probability theory to computing significance levels for rejection of the null hypothesis. The probability levels are usually exact regardless of the shape of the population distributions. But these considerations are only part of the story, for we still are faced with the problem of selecting an appropriate (optimum) statistical test for data analysis. In choosing a statistical test, we must consider the nature of the population from which the sample was drawn, the manner in which the scores were obtained, the kind of measurement (scaling) which was employed to express the scores, and the like.

The problem of selecting an appropriate statistical test may be related to various *research conditions* which stipulate the operational definitions of the variables involved and the questions which can be asked of the data. For this publication, eleven statistical tests have been selected and grouped in terms of six commonly encountered conditions which necessitate a choice between tests. Although these six conditions, and the eleven selected statistical tests, fall short of exhausting the available supply of

Research Conditions	**Appropriate Statistical Tests**	
	I	**II**
ONE SAMPLE	**Binomial Test**	**Chi-Square Test**
TWO RELATED SAMPLES	**Sign Test**	
TWO INDEPENDENT SAMPLES	**Mann-Whitney U Test**	**Median Test** **Chi-Square Test**
MORE THAN TWO RELATED SAMPLES	**Friedman Two-Way Analysis of Variance**	
MORE THAN TWO INDEPENDENT SAMPLES	**Kruskal-Wallis One-Way Analysis of Variance**	**Extension of the Median Test** **Chi-Square Test**
MEASURE OF CORRELATION	**Spearman Rank Correlation Coefficient**	

These eleven tests were selected because of their general usability and directness. Several additional statistical tests are available in the literature. Some of these additional tests are more specialized and may yield more complete information in relation to particular needs.

Figure 8. A Minimum of Statistical Tests To Suit Ordinary Conditions

nonparametric tests (Siegel's book presents twenty-seven), they include sufficient leeway to permit analysis of most instances encountered in administration and teaching. Figure 8 identifies the six research conditions which call for different treatments, and the various statistical tests which are appropriate in each case. We will next turn to a description of what is involved in using each of the eleven tests.

THE ONE SAMPLE CONDITION

The Binomial Test is useful when a single population is conceived as consisting of only two classes. Here we are dealing with a "head or tails" situation where, for example, the probability of a coin's landing heads up is 1/2 and the probability of its landing tails up is also 1/2, so the proportion (P) to be expected for obtaining either a head or a tail when tossing a single coin is also 1/2. But what is to be expected if we toss 20 coins simultaneously? If the null hypothesis is true, we would expect to find about half of the coins landing heads up and half of them landing tails up, when tossed a fairly large number of times.

But when we do research, our statistical interest is in determining the probability of obtaining the observed values—say 6 heads and 14 tails, instead of 10 heads and 10 tails—in a toss of 20 coins. This probability can be computed by applying the formula for the *binomial distribution* which is $(p+q)^n$, where p represents the proportion (frequency) of heads and q the frequency of tails, and n is the number of cases in the sample (both heads and tails). The binomial expansion can be computed to determine how often a sample as unusual as the one observed would come from some hypothesized population wherein we are testing a deliberately introduced independent variable (say psychic powers of the coin-tosser).

However, binomial probabilities for various combinations of two-class events have conveniently been tabled, and from one of the tables we can read directly the probability of most actual occurrences. Use of the table obviates the necessity for computing the binomial expansion. The tables generally let the symbol x equal the smaller number of observed frequencies, and the symbol N equal the total number of frequencies in the sample. For example, referring back to our mention of obtaining 6 heads and 14 tails from a toss of 20 coins, what is the probability of obtaining this exact distribution? Here $x = 6$ and $N = 20$. The tabled probability for these values is .058, or significance is reached at slightly more than the 5 per cent level. To continue with these examples of two categories from the same population, other tabled values which we might have occasion to use are: $N = 8$, $x = 1$, P (probability) $= .035$; $N = 21$, $x = 6$, $P = .039$; $N = 48$, $x = 17$, $P = .030$; $N = 100$, $x = 41$, $P = 044$. As we can see from these illustrations, some binomial tables allow us to work with sample sizes between 3 and 100. The tabled significance values are one-tailed. When a two-tailed test is desired, the significance value is doubled. Thus, when $N = 20$ and $x = 6$, the two-tailed probability is $2(.058) = .116$.

There are natural populations in the school world, the measurements of which may be dichotomized and analyzed through use of the binomial test. Examples of such categories are: boys and girls, Negro and white, stars and isolates, teachers and administrators. For such cases, all possible observations from the population will fall into either one or the other of the two classifications. When the proportion (frequency) of cases in each of the two categories equals 1/2, the prepared tables may be used. When the frequencies in the two categories are not equal to each other, they may be expressed as proportions and substituted in the formula for the sampling distribution of the binomial. But to promote simplicity and parsimony in computational labors, it is generally preferable to select samples and designs which give an equal number of randomly selected cases for each of the two classifications.

The One Sample Chi Square Test is suitable when the researcher is interested in the number of subjects in a single sample which fall into various categories or classes. The number of categories may be two or more. Chi square is a test of the significance of differences—in this case, differences between categories.

As an illustration, suppose that a congenital right-hander tosses a coin 100 times with his left hand and keeps track of the results. He observes that 36 heads and 64 tails appear. We refer to these frequencies as the *observed frequencies.* Next we state the null hypothesis that this distribution does not differ from what we would expect by chance, or 50 heads and 50 tails. These frequencies are called the *expected* (or theoretical) frequencies. The observed and the expected frequencies can be compared through recording them next to each other in what is called a *contingency table,* like this:

Table 1. Comparison of Observed and Expected Frequencies in 100 Coin Tosses

	HEADS	TAILS
OBSERVED	36	64
EXPECTED	50	50

We test our null hypothesis by using the general formula for chi square:

$$\textit{chi square} = \textit{the sum of: } \frac{(\textit{observed frequency} - \textit{expected frequency})^2}{\textit{expected frequency}}$$

The formula requires that we take each observed frequency, subtract from it the corresponding expected frequency, square the difference, and divide the result by the expected frequency. The sum of these operations is chi square; or $36 - 50 = 14^2 = 196/50 = 3.92$, and $64 - 50 = 14^2 = 196/50 = 3.92$, and $3.92 + 3.92 = 7.84$, which is the obtained chi square in this case. We compare this computed value with the values shown in a table of the Distribution of Chi Square. The tabled comparison shows that our observations are significant at better than the 1 per cent level, which equals a chi square value of 6.635. Our computed value is larger than this, but smaller than the 10.827 required for significance at the .001 level.

In this illustration, we determined the level of significance by means of the general formula for chi square. We could have accomplished substantially the same thing through application of the binomial test. Using the same frequencies in our contingency table, we see that we have an N (total frequencies) of 100, and an x (the smaller frequency) of 36. Referring to a table of significance values for the binomial distribution, we find that when $N = 100$ and $x = 36$, $P = .003$.

At this point, we have to be concerned with the tails of the two distributions if we are to compare them. The chi square table is computed for two-tailed tests of the null hypothesis, and the binomial table is computed for one-tailed tests. As we have already seen, we can convert the binomial values to their two-tailed equivalents by doubling the significance values shown in the table. Thus, the tabled value of .003 becomes $2(.003) = .006$. Our obtained chi square significance level was higher than the .01 level and something less than the .001 level. This compares quite favorably with the two-tailed significance level of the binomial test. In both instances, we can be quite confident that these results are different from those produced by chance alone.

THE TWO RELATED SAMPLES CONDITION

The Sign Test is the simplest of all the nonparametric tests. Calculation of the probability of an observed distribution requires nothing more complicated than the simple counting of positive and negative signs, and then referring these values to the appropriate table of probabilities. But simplicity should not be confused, as frequently happens, with superficiality. In the case of the sign test, more accurate synonyms would be "uninvolved" and "elegant." The clear appeal of this uncluttered test is probably responsible for its use in agricultural, medical, and physical research.

In many types of investigation, quantitative measurements cannot be made, or even a complete ranking of the material may not be possible, but cases can be compared in pairs. Thus, experimental and control subjects can be matched in pairs and each subject compared with its control. The only requirement is for each pair to be matched with respect to sex, age, IQ, or other relevant characteristics pertinent to the study being conducted. As we have noted before, one way of accomplishing this is to use each subject as his own control. If observations of each pair can reliably indicate even nominal differences such as "better-worse," "faster-slower," "greater-fewer," etc., the sign test is particularly useful.

Of course there may be instances in which we do not have quantitative measurements (like teacher-made tests) which result in numerical scores for the two members of a matched pair. For example, in both the nonquantitative and the quantitative comparisons, we may be working with a research hypothesis which asserts that the experimental treatment will yield results superior to the control treatment. This, of course, calls for a one-tailed test. The method consists in finding the difference between an observation or score for a member of the experimental group and that of his match in the control group, and then counting over all how many of these differences are positive and how many are negative. When scores are tied for any pair, the difference is zero. In the event of tied scores, the zeros are dropped from the analysis and the N is reduced. The null hypothesis, of course, is that we would expect to find about half of the differences positive and half of them negative. H_0 is rejected if too few differences of one sign occur.

To illustrate the application of the sign test, imagine that we are working in an eighteen-teacher elementary school. Our problem is to improve teachers' attitudes about faculty meetings. Based upon the involvement principle, we hypothesize that if the teachers plan and conduct their own meetings, instead of continuing the customary practice of attending "the principal's meeting," they will reflect an increase in satisfaction with staff meetings. The null hypothesis would be that we could expect about as many dissatisfactions as satisfactions with the new approach. To obtain a measurement of feelings, we construct a simple satisfaction-dissatisfaction scale:

very satisfied	quite satisfied	so-so	somewhat dissatisfied	very dissatisfied
x	x	x	x	x
(5)	**(4)**	**(3)**	**(2)**	**(1)**

Before the change is made, we ask each teacher to express his feelings concerning the customary "principal's meeting." This is done on a prepared sheet of paper, and the sheets are preserved to serve later as control scores. After the change is made, we again ask each teacher to record his feelings by drawing a circle around the x that most nearly represents how he feels. The numerical weights—(5) (4) (3) (2) (1)—do not need to appear on the scale prepared for the teachers, but the use of scale weights helps in recording and interpreting the positive and negative signs.

This procedure gives us 18 matched pairs, and it also gives us a "before-after with control group" design.

Now suppose further that we summarize our data by finding the difference between the first and second scaled response for each teacher, taking care to subtract each time in the same order (say second response minus first response). In line with our research hypothesis, positive and negative differences will indicate increases and decreases in satisfaction. When we record our findings, assume they look like this:

Table 2. Positive, Negative, and Zero Signs for 18 Teachers

Teacher	Sign	Teacher	Sign	Teacher	Sign
1	+	7	−	13	+
2	+	8	+	14	+
3	−	9	−	15	+
4	+	10	+	16	−
5	0	11	0	17	+
6	+	12	+	18	+

By counting these signs, we observe that 2 teachers showed no change between the two treatments (0). They are dropped from analysis, and the N is reduced to 16. There are 4 negative signs, and these are represented by the symbol x. Consulting a table of distribution for the sign test, we observe that for $N = 16$ and $x = 4$, the level of significance is .038, or the null hypothesis can be rejected at slightly more than the 3 per cent level (one-tailed).

It should be apparent that what we were doing is analogous to the binomial test, and that in effect we were seeking the probability of tossing 4 tails (or heads) out of 16 tosses of a coin. But the sign test permits us to deal with a two sample condition, instead of simply a one sample case, and this introduces us to the classical and most often used circumstance in the entire range of research design and statistical analysis.

THE TWO INDEPENDENT SAMPLES CONDITION

The Mann-Whitney U Test is one of the most powerful of the nonparametric tests. "Power" here refers to the relative efficiency of this statistic in rejecting the null hypothesis when that hypothesis is false. And the null hypothesis tested by the Mann-Whitney U Test is that the two independently drawn groups are samples from the same population.

This test is used with independently drawn groups, and the size of the two groups need not be the same (although it may be). It is used to test the difference between the distributions of scores for the two groups, or the difference between two methods or treatments applied to the members of an experimental group and a control group. Scores must be expressed at least in ordinal measurement scales to permit ranking. In fact, this is one of the most frequently used nonparametric "rank tests."

The procedure is quite forthright and uncomplicated. When measurement has been achieved for both the experimental and control groups, the scores of both groups are ranked together in one composite distribution from low to high, assigning the rank of 1 to the score which is algebraically lowest (if negative scores are present), rank 2 to the next lowest, and so forth until all scores receive a rank. Tied scores are assigned the average of the tied ranks. During this joint ranking, it is important to retain the identification of each score as E (experimental group) or C (control group).

The next step requires that the two groups be separated, each score now being assigned the rank determined for it in the composite distribution. The smaller of the two groups is identified by the symbol n_1, and this may be either the experimental or the control group. The larger of the two groups is labeled n_2, and of course may be either the E or C group.

Next, the ranks for each of the two groups are summed. The symbol R_1 is used for the sum of ranks for the n_1 (smaller) group, and the symbol R_2 signifies the sum of ranks for the n_2 (larger) group.

When the values of n_1, n_2, R_1, and R_2 are known, the value of U may be determined. This is done through making substitutions in two formulas. The two formulas yield different values for U. It is the smaller of these two values that is used to determine significance. The symbol U_1 is used to represent the smaller value, and the symbol U_2 denotes the larger value.

The method for determining the significance of the observed value of U_1 depends on the size (number of cases) in the n_2 group. Tables have been prepared for one-tailed and two-tailed tests of $n_2 = 3$ to $n_2 = 20$. For n_2 groups from 3 to 8, exact probabilities are tabled; for n_2 groups between 9 and 20, critical values of U are tabled; and for n_2 groups larger than 20, an approximate test based on the normal curve is available. It has been shown that when n_1 and n_2 are larger than 20, the sampling distribution of U rapidly approaches the normal distribution.

The Median Test will provide information on whether two independent groups have been drawn from populations with the same median. It is applied when individuals are assigned to the two groups (and consequently to two experimental treatments) at random instead of by matching as is the case with the sign test. The null hypothesis states that the two groups are from populations with the same median. The research hypothesis may be that the median of one group is different from that of the other (two-tailed) or that the median of one group is higher than that of the other (one-tailed). The number of individuals in each of the two independent groups does not necessarily have to be the same.

The procedure closely follows the intent of the test. The idea is to compare the numbers of individuals in each group whose scores are above the median score of both groups ranked together and the number of individuals whose scores fall at or below the composite median. If the two groups are samples from populations with the same median, we would expect about half of each group to have scores above the composite median and about half to have scores below this median.

To perform the median test, we first determine the median for the combined scores of both groups. Customarily, this is accomplished by ranking the scores from lowest score to highest score (or vice versa). The median is that value which has as many cases below it as there are cases above it. There are just as many cases with values below the median as there are cases with values above the median. When there is an even number of cases, the median is the value halfway between the two central numbers. For example, if we had an N of 40, and central numbers of 19 and 20, the median would be 19.5. To permit ranking, the scores must be expressed in at least an ordinal scale.

Next we count the number of scores above the joint median for each of the two groups and the number of scores which fall at or below the median for each of the two groups. These values are entered in a 2 × 2 or fourfold contingency table such as the following:

Table 3. 2 × 2 Contingency Table Showing Placement of Frequencies and Totals

	Group I	Group II	Right-hand Margin Totals
Above the Median	a	b	$k\ (a+b)$
At or Below the Median	c	d	$l\ (c+d)$
Bottom Row Margin Totals	$m\ (a+c)$	$n\ (b+d)$	$N\ (k+l)$ or $(m+n)$

If both group I and group II are samples from populations with the same median, we would expect frequencies *a* and *c* to be about equal, and frequencies *b* and *d* to be about equal. The significance of discrepancies from this expectation can be determined by substituting the observed values derived from an actual experiment for the letters *a*, *b*, *c*, and *d* in the contingency table, and then calculating the value of chi square. A formula is available for this purpose which lends itself to machine computation and which avoids the necessity of calculating expected frequencies. This is not the general formula for chi square which was used in connection with the one sample chi square test, and it may be used only in the instance of data which can be set up in a 2 × 2 or fourfold contingency table such as the one just presented for the median test.

If a two-tailed test of the data is desired, we use a table for the Distribution of Chi Square to determine the level of significance. This is done directly because the chi square table of significance values is put up in terms of a two-tailed distribution. If a one-tailed (directional) test is desired, we halve the level of significance shown in the table. For example, according to the table of significance, an obtained value of chi square of 2.706 or more (and less than 3.841) is significant at the .10 level for a two-tailed test. Halving this, we have $.10/2 = .05$ for the significance level of a one-tailed test. It is important to note that *a one-tailed chi square test is permissible only in the instance when a 2 × 2 or fourfold contingency table has been employed as a way to set up the data.* All chi square tests which deal with more than two groups or more than two classifications must follow the general procedures for two-tailed hypotheses and decisions. It is not possible to convert chi square contingency tables larger than 2 × 2 to one-tailed tests.

We will next turn to the chi square two independent samples test and take a brief look at 1) determining expected frequencies; 2) further use of the general formula for calculating the value of chi square; and 3) using the significance table for the Distribution of Chi Square.

The Chi Square Test for Two Independent Samples may be used to determine the significance of differences between two independent groups. This test may be used when the data can be expressed as frequencies (number of cases) in separate classifications (discrete subgroups of the total sample), and all cases can be accounted for among the classifications. The measurement involved may be as simple as nominal scaling, or ordinal and interval scaled data may be used if desired. Numbers of cases in the two groups and in the several classifications do not necessarily have to be the same.

The research hypothesis to be tested usually is that the two groups differ in respect to some characteristic, and therefore in respect to the number of cases which fall into the separate classifications. To test this hypothesis, we count the number of cases from each group which fall into the various classifications, and for each classification, compare the proportion of cases in one group with the proportion of cases in the other group. To do this, we use the general formula for chi square exactly as it was presented for the one sample chi square test.

But now we are faced with two procedures, in addition to the use of the general formula, which were not pointed out in the description of the one sample case. The first procedure concerns a method for determining expected frequencies, and the second procedure concerns the calculation of what is called "degrees of freedom" which is necessary to the use of the Distribution of Chi Square table of significance values.

To illustrate these procedures, a simple example of a research problem involving chi square will be given. We will be using the general chi square formula which directs that we take each observed frequency, subtract from it the corresponding expected frequency, square the difference, and divide the result by the expected frequency. The sum of all these resulting quotients is chi square.

Suppose that we wished to test differences between the 43 sixth grade girls and 52 sixth grade boys ($N = 95$) of one elementary school in regard to their preferences for science, dramatics, and band. This school maintains a schedule of clubs which the children may elect for three meetings per week under faculty guidance. We wish to test whether girls and boys differ with respect to their interests in these three clubs, as expressed by their elective affiliation with one of them. We set up a frequency count of the number of girls selecting each club and the number of boys selecting each club:

Table 4. Girls' and Boys' Choices of Interest Clubs

Club	Number Girls	Number Boys	Right-hand Margin Totals
Science	12	32	44
Dramatics	22	14	36
Band	9	6	15
Bottom Margin Totals	43	52	$N = 95$

The null hypothesis would be that sex is independent of interest; i.e., that the proportion of girls selecting each club is the same as the proportion of boys selecting each club when the total membership of all three clubs is considered. But it is necessary to determine the expected frequency for each count tabulated above.

To do this, for each observed frequency recorded above, we multiply the two totals in the margins common to a particular frequency, and then divide this product by the grand total, *N*, to obtain the expected frequency. For example, the observed frequency of girls selecting the science club is 12. The total in the right-hand margin which is common to 12 is 44, the total in the bottom margin which is common to 12 is 43, and $N = 95$. So we have $44 \times 43 = 1892$, and $1892/95 = 19.9$. Hence, we have an observed frequency of 12 and an expected frequency of 19.9. In the case of the number of boys who selected the band club, we multiply 15×52 (the two totals in the margins common to 6) and divide this product by 95, or $780/95 = 8.2$. We continue to do this same multiplication and division until we have determined an expected frequency for every observed frequency. At this point, it may be pertinent to interrupt with the observation that the expected frequencies amount to a numerical statement of the null hypothesis: these are the frequencies we would have had if there had been no differences.

When we have the observed and the expected frequencies, we can complete the chi square calculation which directs us to subtract each expected frequency (E) from its corresponding observed frequency (O), square the difference, and divide by the expected frequency. The results of this operation for all pairs of observed and expected frequencies are summed, and this is chi square.

When subtracting the expected from the observed frequencies, sometimes the difference is a negative number. But this causes no difficulty because the next step is to square the difference, and negative numbers conveniently disappear. The square of any number, positive or negative, is always positive.

The complete set of steps in the computation of chi square for our problem concerning sixth grade clubs follows:

Table 5. Calculation of Chi Square for Data of Table 4

Clubs	Observed Frequency O	Expected Frequency E	O — E	(O — E)^2	$\frac{(O-E)^2}{E}$
SCIENCE					
Girls	12	19.9	−7.9	62.41	3.14
Boys	32	24.1	7.9	62.41	2.59
DRAMATICS					
Girls	22	16.3	−5.7	32.49	1.99
Boys	14	19.7	5.7	32.49	1.65
BAND					
Girls	9	6.8	2.2	4.84	.71
Boys	6	8.2	−2.2	4.84	.59
TOTALS	95	95	0	(chi square) =	10.67*

(*Significant at the .01 level)

To determine the level of significance for any obtained chi square value, it is necessary to compute the degrees of freedom associated with that value. Chi square is really a family of distributions; that is, there is a different sampling distribution for every value of degrees of freedom. The significance table for the Distribution of Chi Square lists the various numbers of degrees of freedom (usually from 1 to 30) vertically down the left-hand column of the table. When we know the degrees of freedom (*df*), we read horizontally into the table opposite the appropriate listing where we find a row of chi square values. If an observed value of chi square is equal to or greater than the value given in the table for a particular level of significance, at a particular *df*, then H_0 may be rejected at that level of significance.

How is the number of degrees of freedom calculated? In any contingency table composed of *R* rows (classifications) and *C* columns (groups), the number of degrees of freedom is given by $(R-1)\ (C-1)$. In our example, we have three rows (science, dramatics, and band) and two columns (girls and boys). Thus $R=3$ and $C=2$. The number of degrees of freedom is $(3-1)\ (2-1)$, or $2 \times 1 = 2$. We are working at 2 *df* and this enables us to refer to the table for the Distribution of Chi Square to determine the level of significance. In our example, the obtained value of chi square was 10.67. At 2 *df* this value is larger than the tabled value of 9.210 which is significant at the .01 level, but less than the tabled value of 13.815 required for significance at the .001 level. Hence we may decide to reject the null hypothesis at slightly better than the .01 level. We may conclude that there is a difference in girls' interests and boys' interests as expressed by our data.

For a contingency table composed of two rows and two columns, referred to as a 2×2 or fourfold table, the number of degrees of freedom is $(2-1)\ (2-1)$, or $1 \times 1 = 1$. The 2×2 table is a special case always having $df = 1$. You will recall that this set-up is essential to the median test, and that it represents the single instance wherein chi square can be converted to a one-tailed (directional) test.

THE MORE THAN TWO RELATED SAMPLES CONDITION

The Friedman Two-Way Analysis of Variance by Ranks is useful for testing the significance of differences among *three or more* related groups. The null hypothesis to be tested is that three or more samples have been drawn from the same population or populations with the same median. The restriction is that the several groups must each consist of matched individuals, and this of course results in the same number of individuals or groups in each sample. The matching may be achieved by the researcher's selecting three or more groups of subjects matched on certain relevant variables (age, IQ, sex, socio-economic status, etc.) and then randomly assigning each individual or group to one of several experimental situations.

The data are arranged in a two-way table containing *rows* (which correspond to the individuals or groups selected for study) and *columns* (which correspond to the experimental situations). Thus, we work with a table of *N* rows and *C* columns. For example, if one wished to study the differences in learning achieved under four different teaching methods, *N* sets of four matched students might be selected, and one student from each set would be assigned to method I, another from each set to method II, another from each set to method III, and the fourth to method IV. Suppose that we wish to study the scores of six groups under the four teaching methods or situations. Each group contains four matched students, one being assigned at random to each of the four situations. Further suppose that the scores for this study are those given in Table 6.

Table 6. Scores of Six Matched Groups Taught by Four Methods

	TEACHING METHODS			
Group	**I**	**II**	**III**	**IV**
Group 1	22	14	6	19
Group 2	16	11	9	25
Group 3	18	5	10	12
Group 4	29	20	16	21
Group 5	35	33	26	29
Group 6	31	19	17	26

To analyze these data using the two-way analysis of variance by ranks, we rank the observations in each row from 1 to 4, giving the lowest score in each row the rank of 1, the next lowest the rank of 2, and so forth. For example, the four observations in the top row of Table 6 are 22, 14, 6, and 19. These are replaced by ranks 4, 2, 1, and 3. The ranks in each column (situations) are then summed, thus obtaining the data in Table 7.

Table 7. Ranks of Six Groups Taught by Four Methods

	TEACHING METHODS			
Group	**I**	**II**	**III**	**IV**
Group 1	4	2	1	3
Group 2	3	2	1	4
Group 3	4	1	2	3
Group 4	4	2	1	3
Group 5	4	3	1	2
Group 6	4	2	1	3
Rank Totals	23	12	7	18

If the samples (columns) have been drawn from the same population, the ranks in each column will be a random arrangement of the numbers 1, 2, 3, and 4. Under these circumstances, the sum of ranks for columns will tend to be the same, or nearly so. But if the subjects' scores have been dependent on the various teaching methods (situations), then the rank totals will vary from one column to another. If the totals differ significantly, then H_0 may be rejected.

The formula to be applied to the column sums of ranks is Friedman's chi square test. The values given by the Friedman formula are distributed approximately as chi square with $df = C$ (columns) $- 1$. In our example we were working with 4 columns, or $df = 4 - 1 = 3$.

The general table for the Distribution of Chi Square can be used when the number of rows (N) and columns (C) is not too small. For very small N and C, a table of exact probabilities is available for $C = 3$, $N = 2$ to 9; and for $C = 4$, $N = 2$ to 4.

When the investigator is working with individuals not organized into groups, it is feasible to substitute individual subjects for groups in recording row scores. In such an instance, the same sample of N individuals is tested under each one of a number of different experimental conditions, and matching is achieved. Here N equals the total number of subjects (rows) and C equals the number of different experimental conditions (columns). The procedure is the same as that just described for the case of groups randomly assigned to various situations.

Because the Friedman test, by definition, is always concerned with more than one degree of freedom (three or more groups), the sampling distribution of chi square always is analyzed by a two-tailed test. Hence, the research hypothesis is capable of predicting an over-all difference, but not the direction of that difference.

THE MORE THAN TWO INDEPENDENT SAMPLES CONDITION

When three or more groups or situations are to be compared in an experiment, it is helpful to use a statistical test which will indicate whether there is an over-all difference among the groups or situations being studied. If an over-all difference of sufficient magnitude is found to permit rejection of H_0, then any pair of samples may be picked out to test the significance of the difference between them. Otherwise, the two-tailed nature of testing three or more samples approaches scientific value chiefly as a hypothesis generator for more explicit (one-tailed) tests and decisions. We have just considered procedures for testing three or more random samples of equal size which are *matched* according to relevant criteria which may affect the values of the observations. We shall now consider two procedures for testing three or more samples which are *not* matched and which are not of the same size.

The Kruskal-Wallis One-Way Analysis of Variance by Ranks is useful when we wish to test three or more *independent* samples, not of the same size and not matched on the basis of a priori variables. The Kruskal-Wallis technique tests the null hypothesis, as does the Friedman technique, that three or more samples (columns $= C$) are drawn from the same population or from populations with similar averages.

To apply this test, all of the scores for the three or more samples are replaced by ranks in a single series arranged from low score to high score. Of course, at least ordinal measurement is required to permit the ranking of scores. The lowest value is assigned a rank of 1, the next lowest a rank of 2, and so on. The sum of ranks, R_1, R_2, R_3 . . . R_k for each of the C samples, is obtained. When ties occur, the usual convention is adopted of assigning to the tied observations the average of the ranks they would have occupied had there been no ties. N equals the total number of cases in all samples combined.

A statistic, H, is calculated from the data. The formula used results in a value which is distributed as chi square with $df = C - 1$. The general table for the Distribution of Chi Square can be used to determine significance, provided that the sizes of the C samples are not too small. When $C = 3$ and the number of cases in each of the three samples is 5 or fewer, the chi square approximation is not sufficiently close. Kruskal and Wallis have provided a table of exact probabilities for various possible values of n_1, n_2, and n_3 where sizes of the three samples consist of all possible combinations of 5, 4, 3, 2, and 1.

An instance of testing three groups for differences in a single study may be cited from a university course in general public school administration. Class members were to investigate the advantages and disadvantages encountered in X-Y-Z ability grouping. Prior to consulting the literature on the subject, and prior to entering into seminar type discussions, a pretest was administered to determine differences in attitudes on grouping among the three different kinds of administrators enrolled: a) elementary school principals, b) superintendents, and c) secondary school principals. A ten-item attitude inventory was prepared. Each item was to be marked on a nine-point scale from "Strongly Agree" to "Strongly Disagree" with seven points between these two extremes. The scales were weighted toward Strongly Agree $= 9$ and Strongly Disagree $= 1$. Thus a high score over

Table 8. Attitude Scores on X-Y-Z Ability Grouping of Three Samples

Elementary Principals	Superintendents	Secondary Principals
15	15	22
22	19	38
27	36	39
31	47	40
37	48	40
44	61	48
60		63
65		67
		69

Table 9. Attitude Ranks on X-Y-Z Ability Grouping of Three Samples

Elementary Principals	Superintendents	Secondary Principals
1.5	1.5	4.5
4.5	3	10
6	8	11
7	15	12.5
9	16.5	12.5
14	19	16.5
18		20
21		22
		23
$R_1 = 81$ $(n_1 = 8)$	$R_2 = 63$ $(n_2 = 6)$	$R_3 = 132$ $(n_3 = 9)$ $N = 23$

the ten items would indicate relative strength of agreement with ability grouping, and a low score would indicate relative disagreement.

Table 8 presents the scores determined from the completed instruments. The 23 scores were then ranked together from lowest to highest to obtain the ranks shown in Table 9.

From these data, the statistic H was computed by substituting the observed values in the Kruskal-Wallis formula. The computation resulted in $H = 3.174$. Because the number of individuals in each of the three groups was larger than 5, the obtained value of H at $df = 3 - 1 = 2$ was referred directly to a table for the Distribution of Chi Square. This table showed that with two degrees of freedom, an H value equal to or greater than 3.219 would be needed to reach significance at the .20 level. The obtained value of H in this example was less than this (3.174), so significance was reached at just a bit less than the .30 level. The decision in this case was that H_0 could not be rejected. The three groups of school administrators did not differ significantly in regard to attitudes on X-Y-Z ability grouping.

The Extension of the Median Test is a simpler method for testing the differences between three or more independent samples. The data are comprised of C samples of $n_1, n_2, n_3 \ldots n_k$ observations. As before in regard to three or more samples, the null hypothesis is that no difference exists in the medians of the populations from which the samples are drawn.

The median of the combined (jointly ranked) $n_1 + n_2 + n_3 + \ldots n_k$ observations is calculated. Then the samples are separated, and each one is treated exactly as described and illustrated earlier in the case of the Median Test for Two Independent Samples. That is, for each separate sample the number of scores above the joint median is counted, and the number of scores which fall at or below the median is counted. These values are then entered into an R (rows) $\times$ C (columns) contingency table. A chi square test is applied using the general formula for chi square. For this procedure, expected frequencies must be computed. The significance of the obtained value of chi square can be read from the general table of the Distribution of Chi Square at $df = (R - 1)\ (C - 1)$.

Table 10 presents the data for four samples:

Table 10. Scores Recorded for Four Independent Groups

Group I	Group II	Group III	Group IV
3	3	18	14
6	3	18	16
11	4	25	19
14	5	26	22
16	5	29	22
18	8	31	25
21	9		27
33	14		35

The total number of observations is 30. The median is 18. When we arrange these data in a contingency table and compute the expected frequencies, we have the data presented in Table 11.

Table 11. Observed and Expected Frequencies of Scores for Four Groups

	No. Above the Median		No. At or Below the Median			
Groups	Observed	(Expected)	Observed	(Expected)	TOTALS	
I	2	(3.2)	6	(4.8)	8	(8)
II	0	(3.2)	8	(4.8)	8	(8)
III	4	(2.4)	2	(3.6)	6	(6)
IV	6	(3.2)	2	(4.8)	8	(8)
TOTALS	12	(12)	18	(18)	$N = 30$	(30)

With these data, we proceed directly to compute chi square as we did in the case of two independent samples. The obtained value of chi square calculated from this table is 8.28. The number of degrees of freedom is $(4 - 1)\ (2 - 1) = 3$. The tabled value of chi square required for significance at the 5 per cent level is 7.815. The observed value is slightly more than this, and we may conclude that an over-all (two-tailed) difference exists among the four groups.

The Chi-Square Test for More Than Two Independent Samples is a straightforward extension of the chi square test for two independent samples. In general, the test is the same for two independent groups and three or more groups.

The null hypothesis is that the C samples of frequencies or proportions have come from the same or identical populations. This hypothesis may be tested by applying the general formula for chi square where each observed frequency, or proportion, has subtracted from it the corresponding expected or theoretical frequency; the difference is squared and then divided by the expected frequency; and all resulting quotients are summed to yield the value of chi square. As in the chi square test for two independent samples, the data (frequencies or proportions) are entered into an R (rows) $\times$ C (columns) contingency table of discrete categories. These data may be expressed in terms of nominal or ordinal scales.

For example, suppose that we are interested in testing the differences in children's IQ scores, grade level by grade level, in one elementary school. We can set up a contingency table with six columns representing the six grade levels in the school. Then we can categorize the possible range of IQ scores in increases of ten-point intervals, and enter seven such categories in the rows of the table. The blank table might look like this:

Table 12. A Blank Contingency Table for IQ Scores by Grade Levels 1-6

	GRADE LEVELS						
IQ Scores	1	2	3	4	5	6	TOTALS
70 - 79							
80 - 89							
90 - 99							
100 - 109							
110 - 119							
120 - 129							
130 +							
TOTALS							

Here we are working with a very large contingency table which is 6 × 7 in size. The degrees of freedom are $(6 - 1)\ (7 - 1)$, or 30. There are 42 cells in the table, and for the observed frequency recorded in each of them, an expected frequency must be calculated through manipulating the totals in the margins. Obviously, the amount of labor is increased for a problem of this size, but the procedure remains the same as that previously presented for two independent samples.

For this use of chi square, six separate independent samples are being tested simultaneously for an over-all difference between them. The test is two-tailed. It will enable the investigator to determine the significance of any difference detected, but of course will not, by itself, indicate the location of that difference.

THE MEASURE OF CORRELATION CONDITION

Up to this point, we have been concerned chiefly with the effect of a single independent variable. We now consider the problem of determining the degree of simultaneous or concomitant variation of two variables. The data under consideration consist of pairs of measurements. For example, the data may be measures of both height and weight of children, or measures of both intelligence and scholastic performance. The essential feature of the data is that one observation can be paired with another observation for each member of the group. This type of data has two closely related features: correlation and prediction. *Correlation* is concerned with the degree of relation between two variables. *Prediction* is concerned with estimating the extent of one variable from the knowledge of another.

The Spearman Rank Correlation Coefficient was one of the earliest statistics based on ranks and is one of the best known today. It is what is known as a "product-moment" correlation coefficient and is similar to the esteemed (and more difficult) Pearson correlational method. And the Spearman formula enjoys additional advantages of its own. In many situations where ranking methods are used, quantitative measurements are not possible. For example, children may be ranked by teachers on social adjustment, friendliness, or other qualitative factors. In such cases, the data are comprised of sets of ordinal numbers (first, second, third, etc.). Or, if the data are expressed in terms of cardinal numbers (one, two, three, etc.), they may be used directly.

Coefficients of correlation are conventionally defined to take the values +1, 0, and −1. These indicate, in the order given, a perfect positive, neutral, and perfect negative relation between the two variables. And the degrees of relation are always expressed as a decimal fraction of 1.

The Spearman formula directs one to make a list of the subjects. Next to each subject's entry, his rank for the X variable is recorded, and his rank for the Y variable is recorded. The difference between these two ranks is recorded. This difference is symbolized by the letter d. Then each d is squared to obtain a column of d^2. Then the d^2 column is summed. When we know the values of N (number of subjects) and d^2, the coefficient of correlation can be computed. The formula contains these steps:

$$\textit{Coefficient} = \textit{one whole number (1)} - \frac{\textit{six times the sum of } d^2}{\textit{number in sample } (N) \times N^2 - 1}$$

Table 13. Rank Order Correlation Between Reading and Arithmetic

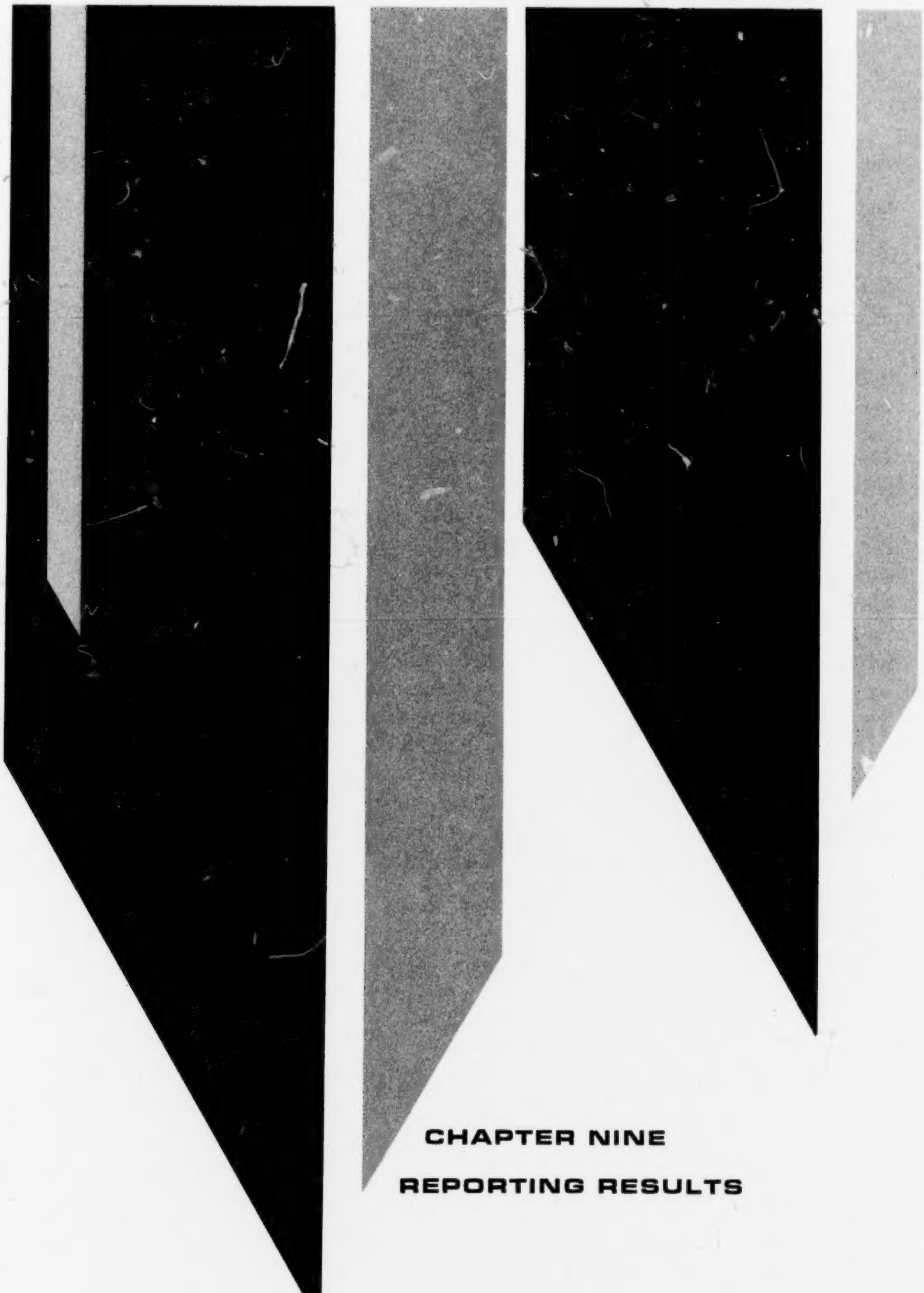
Student	Arithmetic Score	Reading Score	Rank One	Rank Two	d	d^2
A	9.0	9.7	1	3	2	4
B	8.4	11.5	2	1	1	1
C	8.3	9.4	3	4	1	1
D	8.0	10.2	4	2	2	4
E	8.0	7.5	5	10	5	25
F	7.7	8.2	6	7	1	1
G	7.6	9.0	7	5	2	4
H	7.5	8.2	8	7	1	1
I	7.3	8.2	9	7	2	4
J	7.1	7.8	10	9	1	1
					SUM =	46

By way of illustration, one investigator was interested in the strength of the relationship between achievement in reading and achievement in arithmetic. The scores of ten eighth grade students in both areas were selected at random. The results are shown in Table 13.

There are significance tables available for the rank correlation procedure. In this case, the coefficient of correlation computed from the Spearman formula is .722. Reference to a table of significance values shows that for an *N* of 10, this coefficient of correlation is significant at the .05 level. We may conclude that proficiency in reading is significantly associated with proficiency in arithmetic.

As we have seen, the selection of an appropriate statistical test is related to the research conditions which determine the operational definitions of the variable involved and the questions which can be asked of the data. Different research conditions call for different statistical tests. The nature of the population from which the sample was drawn, the way in which the scores were obtained, the kind of scaling which was employed to express the scores—these are among the considerations which must be taken into account in choosing an appropriate test.

While this chapter has not discussed all of the available nonparametric tests, the eleven which have been presented are sufficient to permit analysis under most conditions found in school administration and teaching. The investigator, of course, must exercise wisdom in selecting the test and in applying it to his data. No amount of statistical sophistication can substitute for good judgment.



CHAPTER NINE

REPORTING RESULTS

To say that *all* research projects should eventually be reduced to words and symbols would be to run the risk of appearing pedantic and tedious. Certainly, the custom of prematurely rushing ideas and experiences into writing is schoolish and frequently discouraging. Formal verbalization of ideas and findings which have not been refined through a period of intellectual gestation amounts to little more than the embalming of primitive concepts. No, not all research activities should find their way into written reports. Some should be saved for the dilettante pleasure of simply having fun.

But it would be a mistake to believe that written accounts and reports have little value. The nature of the research process makes it particularly important to record the question being investigated, the steps being taken, the decisions made, and the discoveries encountered. The scientific observer regards his work as "public." He welcomes checks and repetitions of his work as complimentary and as additions to the knowledge he is seeking. He strives to report his work in sufficient detail to enable himself and others to pursue additional related research studies. It is through this kind of report that research becomes cumulative and uncertainty is reduced.

In addition, the researcher generally finds it quite necessary to keep an informal notebook. Such a notebook is used to record immediately data on observations made. Because the human memory is notoriously fallible, the notebook habit is soon nurtured as an indispensable convenience and is seldom regarded as drudgery. Many exploratory and pre-research studies are not taken beyond the notebook phase. But even in these cases, notations may later prove to be extremely suggestive for other studies.

Customarily, a report on completed research, when it appears as an article in a professional journal or when it is filed for the use of colleagues in curriculum improvement programs, follows an outline such as this:

I. Presentation of the Problem with Which the Study Deals

—Review of the literature, if appropriate
—Review of prior studies, if appropriate
—Summary of local issues, if appropriate
—Clear statement of solution being sought

II. Procedures Employed

—Description of population or samples used
—Objectives or hypotheses being tested
—Instrumentation and data collection methods
—Summary tables, when appropriate
—Methods of analyzing data, including statistical tests and decisions

III. Results

- Presentation of findings, including levels of statistical significance
- Logical analyses of major findings

IV. Discussion

- Reference to objectives or hypotheses being tested
- Ideas on extended meanings of the findings for additional studies

V. References

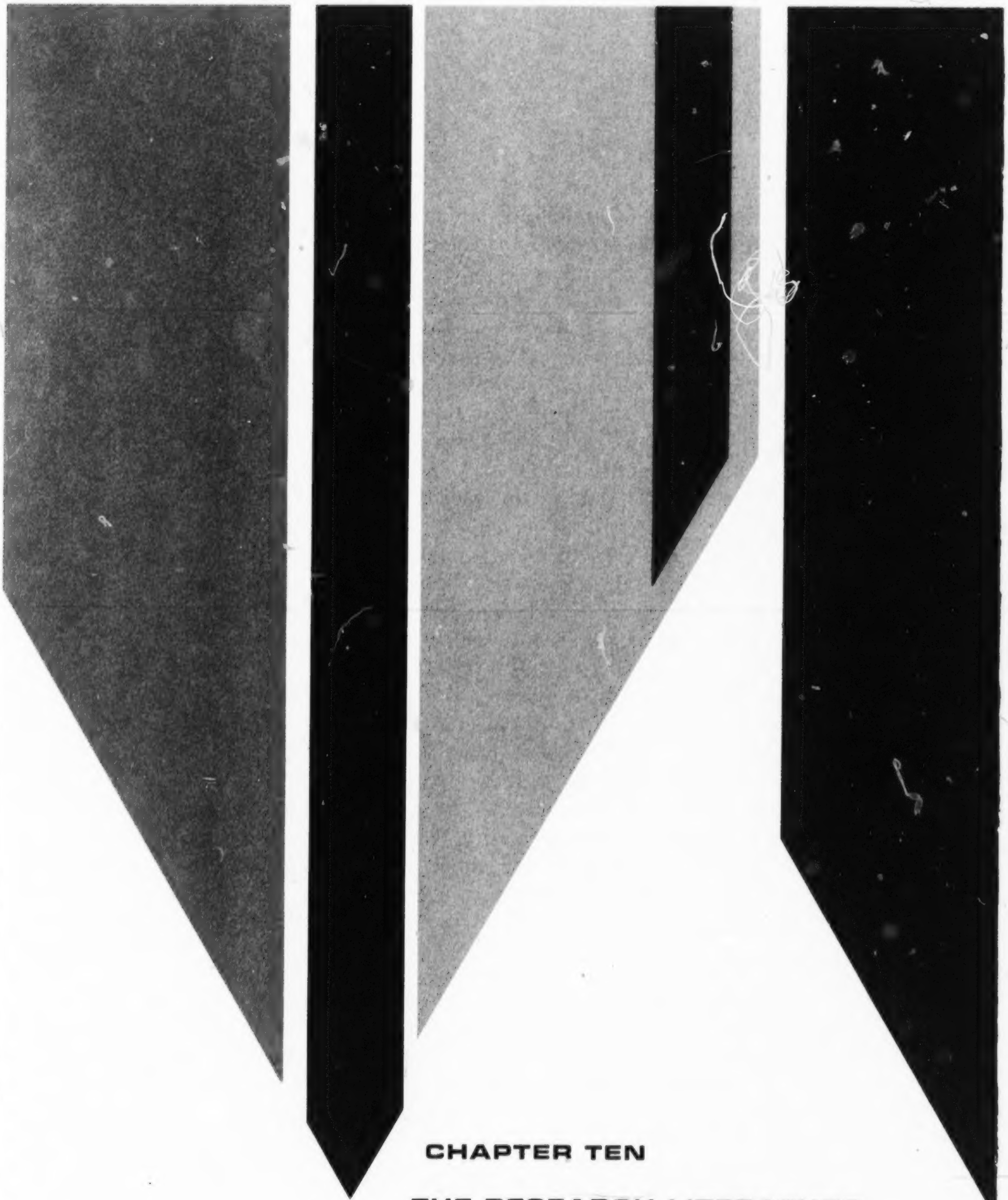
- References in standard form for bibliographical search
- Annotated references, if desired

Although this outline is extremely useful in economically reporting research findings, it unfortunately tends to produce a report which gives unintended and inaccurate impressions of what research actually is. Published reports strongly suggest that research projects have been carried out in a sequential fashion and that they always result in "significant" findings. But the research *process,* in the doing of it, almost never follows the neat sequential pattern suggested in the reports. Furthermore, "failure" to achieve significant findings or to verify predictions made is the usual, not the exceptional, end result of research studies.

The research process would be much simpler logically if the researcher could follow a prescribed sequence of procedure, each step presupposing the completion of the preceding one. In practice, however, it is frequently necessary to take various steps out of order—sometimes to backtrack, other times to compress two or more steps into one, and on other occasions simply to omit one or more steps. Also, the research process involves many additional activities which are rarely mentioned in research reports. For example, methods of data collection may be neatly summarized, skipping the more involved decisions about the kind of data needed and the activities carried out in developing and pre-testing the data collection instruments.

The impression that reported research studies are invariably "successful" may have an even more unfortunate consequence than the impression that the research process follows a prescribed sequence of steps. It may be true that published reports of research studies in education have tended to reveal only the significant and have suppressed the nonsignificant. If so, educational research has been less than half-reported. Perhaps we allow our research activities to become confused with our general workaday obsession with "success" and "failure." These two words, which have taken on a moral overtone, are out of place in reference to research. Nowhere is this more apparent than in relation to the research activities of American industry. In many industries, research projects that result in new and profitable discoveries are distinctly in the minority. The majority of the projects do not work out as intended, but they are carefully preserved because knowing where the potential blind alleys are is greatly important. In a very real sense, no carefully constructed research project can be worthless, regardless of its results. Every project can add useful information to our knowledge and thinking. And the nonsignificant studies can make a needed contribution if they are presented in written form so that we and others can think about them as we plan for similar studies.

It is quite reasonable to expect that each individual school and each individual school system will one day keep a library of research reports completed by its own staff. As more and more school systems allocate an annual budget to support local research, it will be increasingly expected that the outcomes be visible. The NEA's Project on Instruction report, *Schools for the Sixties*, makes this recommendation: "The National Committee has recommended that each school system allocate a definite and adequate proportion—at least one per cent—of its annual operating budget for the support of research, planning and development."[15] When this is done regularly and when research activities become a budgetary line item, school people, board members, parents, and colleagues will develop a regular hunger for local reports which indicate scientific growth in school development and staff improvement.



CHAPTER TEN

THE RESEARCH LITERATURE

The literature on teaching which is found in college textbooks and many professional journals for the most part does not include reports on research. The conclusions of research studies and their implications for practice may be referred to, but they are mentioned in quite general treatments. Such general discussions are useful in undertaking to build a background for a given subject, but they lack the detail and specificity of original reports. Particularly the administrator who wishes to build a background in current research subjects will find himself in need of different and more specialized sources of information from those provided by the customary literature.

It is entirely reasonable to assume that few administrators and teachers subscribe to specialized research journals or have knowledge of their sources. This should not be surprising because traditional programs of teacher preparation, and much current in-service literature, are not focused on the understanding and use of original research reports. But for the person who would become involved and develop skill in conducting his own research, acquaintance with the small list of journals most likely to contain papers on the subjects of his interests will be most rewarding.

It is very important to keep a steady stream of new ideas flowing into a research project. Often these new ideas come from browsing in specific reports of what other researchers are doing, how they state their problems, treat their data, arrive at conclusions, and the like. Becoming overwhelmed with reports on what is being done should be guarded against, and a search of the literature should not be carried too far or it will go on forever and serve as a complete bar to action. On the other hand, even the wise selection of pertinent literature is impossible unless the investigator has knowledge of the structure and sources of research literature. The effort to acquire this knowledge will repay itself handsomely in saving time spent in the library and in selecting journals for private subscription.

In regard to the literature, there are two important and feasible goals for the researcher. The first of these is to find whether the problem area of a proposed research project is being investigated elsewhere, the names of principal investigators in the area, and the kind of investigations being pursued. The second goal is to acquire reassurance in broadly questioning the practices of schools and teaching. It should be borne in mind that findings coming from research projects in education and in the behavioral sciences in general are seldom conclusive or final. The extent to which the generalizations of even the most carefully conducted research project can be applied to a specific classroom per se is probably very small. The foregoing two statements suggest that it is preferable for the investigator in his own situation to "try" experimentally and check out others' research findings before he adopts them. This sort of replication is one of the best and most needed contributions to the broad field of research on teaching. It also represents one of the finest uses of the research literature. Knowing where to locate and how to use the available material can be a desirable addition to the educator's background.

BOOKS

Books and textbooks provide basic material and a general level of information. They provide one of the first approaches to a new specialization in order to gain a feeling for the whole field. There is no simple way of finding all the books on a given subject. The most immediate place to look is the subject index of a good library.

There are a few books which attempt to organize material on methods, research areas, and techniques of searching which seem generally useful. The titles which are cited as references in this volume are not repeated here, although they, too, may be considered part of this list.

Carter Alexander and Arvid J. Burke. **How to Locate Educational Information and Data.** Fourth edition. New York: Bureau of Publications, Teachers College, Columbia University, 1958.

The book a) brings the user up to date on hundreds of sources for locating educational information and data, and b) makes revisions in methods and references contained in earlier editions which save time and energy for users. Included in the book are two main headings: 1) Basic Techniques of Library Utilization which includes chapters on locating books and periodicals, using the indexes, library reading, and note taking; and 2) Special Application of Library Utilization Techniques which includes chapters on book evaluation, government documents, reference books, quotations, statistics, biographical information, etc.

O. K. Buros, editor. **The Fifth Mental Measurements Yearbook.** Highland Park, New Jersey: The Gryphon Press, 1959.

Lists 1,957 commercially available tests published during the years 1952-1958. Also included are over 6,000 references on the construction, use, and limitations of specific tests. There are tests of bookkeeping, Latin achievement, English progress, nursing, sensory acuity, coordination, sales aptitude, handwriting, etiquette, sex knowledge, health, religion, honesty, and many other skills and aptitudes.

Stephen M. Corey. **Action Research to Improve School Practices.** New York: Bureau of Publications, Teachers College, Columbia University, 1953.

Addressed to practitioners in education and stresses cooperative research.

It contains a stimulating chapter on research in education, several interesting accounts of action research projects, and an excellent 80 item bibliography.

N. L. Gage, editor. **Handbook of Research on Teaching.** A Project of the American Educational Research Association, NEA. Chicago: Rand McNally Company, 1963.

This book has been called the most important contribution to educational research methodology yet made. It contains 23 sections written by 31 authors, each of whom is a specialist in his particular field. The book is organized into four parts: Part I, "Theoretical Orientations;" Part II, "Methodologies in Research on Teaching;" Part III, "Major Variables and Areas of Research on Teaching;" and Part IV, "Research on Teaching Various Grade Levels and Subject Matters." The 1,218 pages of the book present a sufficiently comprehensive range of basic material to serve the interests and needs of most people interested in research on teaching.

Jerome J. Hausman, editor. **Research in Art Education.** Ninth Yearbook. Washington, D.C.: The National Art Education Association, NEA, 1959.

A compilation of completed research studies in art education. Five main divisions—philosophical-psychological research, research and creative behavior, research into teaching process, surveys and descriptive research, and research into problems of teaching handicapped and exceptional children—are carefully explored by outstanding authorities in the art field at both the high school and university levels.

C. W. Hunnicutt and William J. Iverson, editors **Research in the Three R's.** New York. Harper & Brothers, 1958.

Contains studies of varying levels of scientific quality dealing with reading, writing, and arithmetic. The works included were chosen for their significance, influence, and importance in the field. Authors of the studies are well-known people with special interests in these areas.

Gardner Lindzey, editor. **Handbook of Social Psychology.** Vol. I, "Theory and Methods." Cambridge: Addison-Wesley Publishing Company, Inc., 1954.

There are nine chapters devoted to research methods. Like handbooks in many of the sciences, this one is large, technical, and informative.

Phi Delta Kappa. **Symposium on Educational Research.** Bloomington, Indiana: the Fraternity.

This series of volumes includes formal papers and subsequent discussions comprising two-day symposia considering different aspects of educational research. The first book (1960) dealt with the current stage of development in educational research; the second (1961), with research design and analysis.

Copies of these books are published each year and may be obtained by writing to Phi Delta Kappa, Inc., 8th and Union, Bloomington, Indiana.

Phi Delta Kappa. **Research Studies in Education.** Bloomington, Indiana: the Fraternity.

These studies are subject-author indexes and research methods bibliographies. Dissertations, reports, and field studies under way or completed at colleges and universities in the U.S. and Canada are classified by subject under general headings.

Available in most university and college libraries. New editions appear yearly.

ENCYCLOPEDIAS

The next level of specialization for most of the sciences consists of the various encyclopedias which are extremely useful for acquiring a comprehensive, yet detailed, view of a given field. Only one encyclopedia centering on educational research, which is sponsored by The American Educational Research Association, is included in this list. It provides a convenient one-volume source book of unquestioned authority, providing researchers with recent developments in their areas of special interest.

Robert Ebel, editor. **Encyclopedia of Educational Research.** Fourth Edition. The Macmillan Company. New York. 1969.

The encyclopedia presents a critical evaluation, synthesis, and interpretation of all the pertinent research—early as well as recent—on a variety of subjects spread over the entire range of educational topics. Four editions of the volume have been published: 1941, 1950, 1960, and 1969. The American Educational Research Association has been responsible for the publication, with two different educational leaders serving in the editorial capacity. Each new volume updates the material found in the other encyclopedias. At the end of each topic, there is an exhaustive bibliography of both books and periodicals which are concerned with the specific topics treated.

LITERATURE GUIDES

Beyond encyclopedias, one comes to the general guides for selection of literature. The guides generally classify titles listed and sometimes include reviews and criticisms of the content. Catalogues sometimes are descriptive and frequently are a valuable help in classifying available literature.

Booklist and Subscription Books Bulletin. American Library Association. Chicago.

Formerly the *Subscription Books Bulletin*, the periodical was combined with another publication of the American Library Association, *Booklist*, in 1956. The publication is used mainly by librarians in evaluating new books. Up to 1956, the periodical gave critical reviews of various reference books, e.g., encyclopedias, indexes, etc. In this new publication, other books are also included in the critical reviews.

Published twice a month September through July and once in August. Single copies may be obtained by writing to the American Library Association, 150 E. Huron St., Chicago, Illinois.

Cumulative Book Index. The H. W. Wilson Company. New York.

A world list of books in the English language. Information about any book is available through three sources in one alphabet: author, title of the book, and subject. For a book for which the title is known, information may be obtained through the publisher's name. Information is arranged under names of persons by:

1. Works which the person authored
2. Works of which he is joint author, editor, or translator
3. Works about the person.

Under names of places the order is:

1. Official publications by departments or bureaus
2. Works about the place
3. Societies, institutions.

Published every four or five years, depending on the magnitude of the number of entries to be indexed. It was first published in 1932, and the current edition is a 1959 publication.

Education Index. The H. W. Wilson Company. New York.

A cumulative subject title index to a selected list of educational periodicals, books, and pamphlets. Near the beginning of each volume are listed the various periodicals from which material has been indexed. The first volume was published in 1932 and contained material written from 1929-1932 (3 years). At present, a volume is accumulated every three years. Small paper-bound copies of current months are available so that the reader may find the latest literature if he wishes to do so.

NEA Catalogue of Publications. National Education Association. Washington, D.C.

Lists publications under the headings of the NEA units responsible for production. Incorporated also are alphabetical and subject indexes. In each catalogue, the publications are listed first by number and then are referred to by that number afterward in the indexes. Entries in green print are new issues. There is a research division of NEA, and its publications are included in the listing.

The recent catalogue may be obtained by writing to the National Education Association, 1201 Sixteenth St., N.W., Washington, D.C.

Reader's Guide to Periodical Literature. H. W. Wilson Company. New York, 1957.

An author and subject index of articles taken from many American periodicals. These publications are listed in the front of each issue. Provides a well-balanced selection of popular, nontechnical magazines representing all the important scientific, technical, and subject fields. Current issues of *Reader's Guide* are paper-backed publications and are produced every two or three months by the H. W. Wilson Company, 950 University Avenue, New York, New York. Available only in libraries.

ABSTRACTING AND INDEXING JOURNALS

These journals are one of the chief sources for the investigator who is seeking information on specific subjects. They also provide one time-saving method of keeping up with current research emphases.

Dissertation Abstracts. University Microfilms. Ann Arbor, Michigan.

A monthly compilation of abstracts of doctoral dissertations submitted by more than 115 cooperating institutions. Each issue consists of a principal section, an author index, and a subject index. The principal section contains the abstracts, arranged under the subject categories assigned by author. The alphabetical list of categories is given in the table of contents. Author and subject indexes are cumulated annually and include "see also" references to more specific or closely related subjects. Also, a xerox copy of any dissertation that has been completed in the United States can be secured from University Microfilms.

For a subscription or for a xerox copy of a dissertation, write to University Microfilms, Xerox Corporation, 300 Zeeb Road, Ann Arbor, Michigan.

Education Abstracts. UNESCO Publications Center. New York, New York.

Each issue of the publication deals with one topic only. Abstracts are listed by country and are sometimes written in the native language. Many of the issues center on education in other countries, curriculum, and educational methods.

Published every month except July and August. Single copies may be obtained by writing to the UNESCO Publications Center, 801 Third Avenue, New York, New York.

Psychological Abstracts. American Psychological Association, Inc. Washington, D.C.

Abtracts of various studies being carried on in all major areas of psychology are reported. Problems dealt with include the nervous system, learning, psychoanalysis, personality, crime and delinquency, mental testing, and others. Each volume includes one calendar year, and is indexed by subject and author. Monthly editions are also available for current studies.

Published monthly with two issues in December. Copies may be obtained by writing to: The American Psychological Association, 1333 Sixteenth Street, N.W., Washington, D. C.

Sociological Abstracts. Leo P. Chall, editor. New York, New York.

The purpose of the periodical is to present abstracts regarding, in order of priority, 1) sociological periodicals and books; 2) periodicals and books bearing on sociology; and 3) nonsociological periodicals containing significant, applicable material. Abstracts are listed under subject areas by author's name.

The periodical was first published in 1952 and is now issued five times a year in January, April, July, October, and November. Single isuses are available by writing to the Editorial Office, c/o Leo P. Chall, 225 West 86th Street, New York, New York.

REVIEW JOURNALS

Review journals are indispensable in enabling researchers to have some idea of what is being accomplished along various lines of investigation. These sources generally provide more complete information than the abstracting journals through discussing reports in terms of their theoretical settings.

Psychological Review. American Psychological Association, Inc. Washington, D.C.

Devoted to articles of theoretical significance to any area of scientific endeavor in psychology. Ordinarily the periodical does not report original research but makes exception where this research is included in a major theoretical article which attempts to integrate several related original studies.

Published bimonthly and available by writing to the American Psychological Association, Inc., 1333 Sixteenth Street, N.W., Washington, D.C.

Review of Educational Research. American Educational Research Association, NEA. Washington, D. C.

The purpose of the *Review* is to report major research findings during a designated period, organized by areas of interest. Significant studies are identified, summarized, and critically analyzed. The more active fields of educational research are reviewed every three years; the less active fields are included in alternate cycles.

Published in February, April, June, October, and December. This journal is a "must" for researchers in education. Single copies may be obtained by writing to the American Educational Research Association, National Education Association, 1201 Sixteenth Street, N.W., Washington, D.C.

RESEARCH JOURNALS

These journals are the primary source of complete and full-flavored material in "reported" research projects. Commonly, they contain descriptions and discussions of techniques, designs, analyses, and sampling methods used. They tend not to be unduly compressed, and contain sufficient detail to make possible critical reading and replications by interested people. These journals provide the richest type of primary acquaintance with the literature on research in action.

The journals listed here are grouped in three major sub-classifications: education, psychology, and sociology.

Education

American Educational Research Journal. American Educational Research Association, NEA. Washington, D.C.

This journal made its debut in January 1964. It publishes original reports of experimental and theoretical studies in education. Reports included are complete, containing information on purposes of the investigation, design and procedure, description of the sample studied, results, discusssion, tables and charts, and conclusions. Reviews of books on research are included in each issue.

Published four times a year. Single copies may be obtained by writing to the American Educational Research Association, 1201 Sixteenth Street, N.W., Washington, D.C.

California Journal of Educational Research. California Teachers Association. Burlingame, California.

Reports on: a) city and county school research pertaining to curriculum, guidance and counseling, evaluation, supervision, and finance; b) digests of theses and dissertations that have practical application; and c) studies that present novel, but tested, approaches to the solution of educational problems. Certain articles may also be found in *Psychological Abstracts.*

Published five times a year: January, March, May, September, and November. Single copies are available upon request from *California Journal of Educational Research,* 1705 Murchison Drive, Burlingame, California.

Canadian Education and Research Digest. Canadian Education Association. Toronto, Ontario.

Includes articles on various aspects of the Canadian educational system. Some of the writing is in French. Also included in some issues are abstracts of staff studies and summaries of graduate theses in education.

This publication is issued quarterly by the Canadian Education Association, 559 Jarvis Street, Toronto 5, Ontario. Subscriptions to the *Digest* may be obtained by writing to the address above.

Educational and Psychological Measurement. G. Frederic Kuder, Durham, North Carolina.

Devoted to reporting: 1) discussions of problems in the field of the measurement of individual differences; 2) reports of research on the development and use of tests and measurements in education, industry, and government; 3) descriptions of testing programs being used for various purposes; and 4) miscellaneous notes pertinent to the measurement field. Book reviews at the end of each issue provide current sources of information for further study.

Issued quarterly; one complete volume per calendar year. For single copies write to *Educational and Psychological Measurement,* Box 6907, College Station, Durham, North Carolina.

Educational Research Bulletin. Bureau of Educational Research and Service, Ohio State University. Columbus, Ohio.

Features articles on current areas of interest in educational research and reviews books that are new in the various interest areas.

Published monthly except during June, July, and August. Copies may be obtained from: Publications Office, Ohio State University, 242 W. 18th Avenue, Columbus, Ohio.

Illinois School Research. Illinois ASCD. Springfield, Illinois.

Contains reports of research conducted in schools on real school problems. Emphasis is on techniques used, methods employed, and major outcomes. Authors include public school personnel, college professors, specialists, and local unit research directors.

Each magazine issue carries interesting articles describing research, experimentation, and other projects in Illinois schools. Contributions to the magazine are welcome and sought.

This journal is published three times a year and distributed to members of the Illinois ASCD as part of their annual membership. Subscriptions available at $3 a year. Order from Administrative Relations Director, Illinois Education Association, 100 East Edwards Street, Springfield, Illinois.

Journal of Educational Research. Dembar Publications, Inc. Madison, Wis.

Reports significant research being carried on in a wide variety of areas. It attempts to be inclusive by stating very clearly the problem, method of procedure, summary conclusions, and implications. Each article is a whole study, not a synopsis or simply a review. There are certain issues which are devoted entirely to one subject, and annotated bibliographies are included.

Published monthly nine times a year. Single copies may be obtained by writing to Dembar Publications, Inc., Box 1605, Madison, Wisconsin.

Journal of Experimental Education. Dembar Publications, Inc. Madison, Wisconsin.

This is a periodical report of scientific investigations relating to child development, curriculum, learning, teaching supervision, measurements, statistics, and experimental techniques. Whole research projects are reported and a selected bibliography is included with each study.

Published quarterly in September, December, March, and June. Single copies are available by writing to Dembar Publications, Inc., Box 1605, Madison, Wisconsin.

NEA Research Bulletin. National Education Association. Washington, D.C.

Research reported is of a cumulative (survey) type. Information on a subject is accumulated and the figures are reported. There is little effort to interpret data in terms of the conclusions and implications involved. Information that has been brought together can be used, however, to view trends in the various areas studied.

Published in March, May, October, and December. Single copies may be obtained by writing to Publications-Sales Section, NEA, 1201 Sixteenth St., N.W., Washington, D.C.

Psychology

American Journal of Psychology. University of Texas. Austin, Texas.

Reports in a sophisticated way research studies in the field of experimental psychology. Five departments are represented in every issue: articles, minor communications, descriptions of apparatus, notes and discussions, and reviews of books. Studies are reported in their entirety and include purpose, procedure, subjects, and summary and conclusions.

The *American Journal of Psychology* is published quarterly, on the 15th day of March, June, September, and December. Back numbers and volumes may be obtained from the Business Manager, Department of Psychology, Mezes Hall, University of Texas, Austin, Texas.

British Journal of Educational Psychology. British Psychological Society. London, England.

Reports research in a rigorous, traditional way. Introduction, procedures, results, and discussion are the rule for every paper. The topics dealt with are very specific in nature and require a sophisticated knowledge of statistics and psychology for understanding. This publication is issued three times a year, in February, June, and November, and may be obtained by writing to the Manager, B.J.E.P. Department, Methuen and Co., Ltd., 36 Essex Street, Strand, London, W.C. 2, England.

Child Development. The Society for Research in Child Development, Inc. Purdue University. Lafayette, Indiana.

The publication is devoted to basic research and theory concerning the growth and development of the child. Each issue is composed of a number of completed research projects, reported fully by men of stature in the field. An attempt is made in each study to report background work, procedure, composition of the sample, summary, conclusions and implications, and discussion of the results.

Published quarterly in March, June, September, and December. Single copies may be obtained by writing: Child Development Publications, Purdue University, Lafayette, Indiana.

Journal of Abnormal and Social Psychology. American Psychological Association, Inc. Washington, D.C.

Emphasis is placed on basic research and theory rather than on the techniques and arts of practice. Papers in abnormal psychology contribute to the fundamental knowledge of the pathology, dynamics, and development of personality or individual behavior. Social psychological papers contribute to the basic knowledge of interpersonal relations and of group influences on the pathology, dynamics, and development of individual behavior.

Issued bimonthly in January, March, May, July, September, and November; two volumes per year. Single copies may be obtained by writing to the American Psychological Association, Inc., 1333 Sixteenth St., N.W., Washington, D.C.

Journal of Educational Psychology. American Psychological Association. Washington, D.C.

Publishes original investigations of problems of learning, teaching, and the psychological development of the individual. It contains articles relating to the development of interests, attitudes, personality, social relations, and vocational orientation with respect to school adjustment. The impact of the school program upon development and adjustment, emotion, motivation and character, mental development, and methods is also explored. Studies published deal with all levels of education —pre-school, college, and adult education, and with all age groups.

Copies (published bimonthly) may be obtained by writing to the American

Psychological Association, 1333 Sixteenth St., N.W., Washington, D.C.

Journal of Experimental Psychology. American Psychological Association, Inc. Washington, D.C.

Reports original experimental investigations which are intended to contribute toward the development of psychology as an experimental science. Studies using normal human beings as subjects are emphasized, except when abnormal or animal subjects lend themselves to extension of behavior theory.

Published monthly and may be obtained by writing to the American Psychological Association, 1333 Sixteenth St., N.W., Washington, D.C.

Journal of Genetic Psychology. The Journal Press. Provincetown, Massachusetts.

Devoted to the study of child behavior, animal behavior, and comparative psychology. Research studies are sophisticated and rigorously reported. Articles are footnoted and contain references which may be used to investigate the topic more carefully.

A quarterly publication, it is issued in March, June, September, and December. Single issues are available at the Journal Press, 2 Commercial Street, Provincetown, Massachusetts.

Journal of Personality. Duke University Press. Durham, North Carolina.

Current stress is placed on experimental studies of behavior dynamics and character structure, personality related consistencies in learning and perception, and the development of personality in its cultural context. Most of the contributions are empirical in character.

Published quarterly in March, June, September, and December. Back single issues are available by writing to the Duke University Press, Box 6697, College Station, Durham, North Carolina.

Sociology

Journal of Educational Sociology. The Payne Educational Sociology Foundation. New York University, New York, New York.

Articles on research in the field of educational sociology are written by authorities in their respective specialties.

Published monthly from September through May. Copies may be obtained by writing to: Publication and Business Office, New York University, Washington Square, New York, New York.

Journal of Social Psychology. The Journal Press. Provincetown, Massachusetts.

Reports formal studies in political, racial, and differential psychology. The research is of a formal nature and is reported in a traditional, sophisticated way. Each article is footnoted and is followed by a list of references which may be used for further study in various interest areas.

Published quarterly in February, May, August, and November. Single copies may be ordered from: the Journal Press, 2 Commercial Street, Provincetown, Massachusetts.

Sociometry. The American Sociological Society. New York, New York.

Emphasis is placed on reporting research dealing with measurement of social behavior. Generalizations are derived from the investigation of the processes and products of social interaction at the interpersonal, intrapersonal, intergroup, and intragroup levels of consideration.

Published four times a year in March, June, September, and December. Available in single copies by writing to The American Sociological Society, New York University, Washington Square, New York 3, New York. Back issues dated 1937-1955 should be ordered from Beacon House, Beacon, New York.

GENERAL JOURNALS ALSO CONTAINING RESEARCH PROJECTS

There are many journals which carry articles in the area of current research studies, research methodology, and reviews of publications on research. Although such articles may not represent a primary focus of the journals, the researcher will do well to include such publications in his list of places to look.

Adult Education (American). Adult Education Association of the United States of America. Norman, Oklahoma.

Deals with current thinking in the field and reports some research being carried on in various areas of adult education. Both articles and research topics are reported in a way which is conducive to easy reading and full understanding.

Published quarterly—autumn, winter, spring, and summer. Single copies are available by writing to: Adult Education Association of the U.S.A., 743 North Wabash Avenue, Chicago, Illinois.

Adult Education (British). National Institute of Adult Education. London, England.

Both a record of activities and an open forum for the discussion of matters relating to adult education. It contains articles of a nonresearch oriented character. Book reviews, pamphlets and reports, and notes are topics included in the format in addition to articles which are written by men from many different fields of interest.

Published quarterly in June, September, December, and March. Single copies may be obtained by writing to the National Institute of Adult Education, 35 Queen Anne St., London W. 1, England.

British Journal of Educational Studies. Faber and Faber, Ltd. London, England.

Presents articles dealing with various aspects of the English educational system, current news in the field of education, correspondence, book reviews, and short notices of new books. Although research is not reported, the articles, written by outstanding university professors and lecturers, are enlightening with respect to the interests and concerns of English education. The writing is clear and fully footnoted.

Published twice yearly in May and November. Single copies are available by writing to Faber and Faber, Ltd., 24 Russell Square, London, W.C. 1, England.

Child Study. Child Study Association of America. New York, New York.

Centered on parent education, so is relatively free from the rigorous type of scientific research present in other journals. The publication deals with a central topic in each issue. Articles are written mainly by medical doctors but also by parents and other people interested in children.

Published quarterly in the summer, fall, winter, and spring. Single copies are available by writing to the Child Study Association of America, 132 East 74th Street, New York, New York.

Childhood Education. Association for Childhood Education International. Washington, D.C.

Presents articles by qualified people, dealing with elementary school age youngsters. Although research topics are not reported as such, they are quoted and cited in the articles. Current books and pamphlets for both children and adults are reviewed.

Published monthly September through May. May be obtained in single issues by writing to *Childhood Education*, 3615 Wisconsin Avenue, N.W., Washington, D.C.

Educational Leadership. Association for Supervision and Curriculum Development, NEA. Washington, D.C.

Centered on a general curriculum theme with features on specific aspects. Authors of the articles represent state departments of education, public schools, and universities. Incorporated into each publication is a section entitled "Research in Review," which summarizes some of the current experiments and research projects being carried on in the area of the general theme.

Copies of *Educational Leadership* are published monthly, October through May, and may be obtained by writing to the Association for Supervision and Curriculum Development, NEA, 1201 Sixteenth St., N.W., Washington, D.C. D.C.

The Educational Record. American Council on Education. Washington, D.C.

Presents articles and some research concerned with various aspects of higher education. Written by prominent men of the nation's colleges and universities, material deals with a wide variety of problems. Research is reported in a condensed, easy to read style, with the emphasis placed on findings and conclusions rather than on procedures and history.

Issued quarterly in January, April, July, and October. Single copies may be ordered through the American Council on Education, 1785 Massachusetts Avenue, N.W., Washington, D.C.

Elementary School Journal. Department of Education of The University of Chicago. Chicago, Illinois.

Presents current articles and research topics in the field of elementary education by authors representative of both the public schools and the colleges and universities. Emphasis is placed on the nonstatistical, nontechnical type of reporting in most cases. Annotated selective references on general topics are incorporated into the format of each publication.

Published monthly October through May. May be obtained by addressing a request for single copies to the University of Chicago Press, 5750 Ellis Avenue, Chicago, Illinois.

Exceptional Children. Council for Exceptional Children, NEA. Washington, D.C.

Composed of: a) empirically grounded research studies of current interest and b) articles on various problems regarding exceptional children that report progress or state opinions. A section of every issue is devoted to reviewing new books in the field.

Issued monthly, September to May. Single copies are available by writing to the Council for Exceptional Children, NEA, 1201 Sixteenth Street, N.W., Washington, D.C.

Journal of Education. Boston University, School of Education. Boston, Massachusetts.

Since December 1955, this magazine has considered a single topic and has featured articles by only one or two authors. The subjects picked concern both elementary and secondary schools. Some of the issues are devoted entirely to reporting research studies, while others simply state what has been learned from the literature and research.

Published quarterly in October, December, February, and April. Single copies may be obtained by writing to *Journal of Education*, 332 Bay State Road, Boston, Massachusetts.

Journal of Negro Education. Howard University Press. Washington, D.C.

The purpose of this journal is threefold: first, to stimulate the collection and facilitate the dissemination of facts about the education of Negroes; second, to present discussions involving critical appraisals of the proposals and practices relating to the education of Negroes; and third, to stimulate and sponsor investigations of problems incident to the education of Negroes.

Published quarterly in winter, spring, summer, and fall. Single numbers are available by writing to the *Journal of Negro Education*, Howard University, Washington 1, D.C.

Journal of Teacher Education. National Commission on Teacher Education and Professional Standards, NEA. Washington, D.C.

Composed of articles and research studies dealing with various aspects of teacher education. Some of the research topics are reported in detail and include conclusions and implications. Other studies and articles merely cite research done in that area or state an educational viewpoint. "With the Researchers," a regular feature by the research editor, draws attention to significant problems or findings in the field which need recognition.

Published in March, June, September, and December. Single issues may be obtained by writing to the *Journal of Teacher Education*, 1201 Sixteenth Street, N.W., Washington, D.C.

The National Elementary Principal. National Association of Elementary School Principals, Arlington, Virginia.

Articles in each issue center on a single topic in elementary education, such as handwriting, grouping, art, etc.

Articles are published every year which incorporate much of the research concerning the elementary school principalship. Research is cited in the body of articles and in the bibliographies which are listed at the end of many articles.

Published in September, November, January, February, April, and May. Single copies may be obtained by writing to the *National Elementary Principal,* 1201 Sixteenth Street, N.W., Washington, D.C.

Phi Delta Kappan. Phi Delta Kappa. Bloomington, Indiana.

This is the official journal of Phi Delta Kappa, professional educational fraternity. Research articles, reviews, and news alert the professional educator and interested layman to significant developments in education.

The *Kappan* is published nine times a year, from October through June. Subscriptions may be obtained by writing to Phi Delta Kappa, Inc., 8th and Union, Bloomington, Indiana.

School and Society. Society for the Advancement of Education. Lancaster, Pennsylvania.

Research studies are incorporated into the periodical on a limited basis. All articles are very short, easy to read, and run the full range of educational topics.

The magazine, in its fortieth year of publication, is published biweekly. Single copies may be obtained from Business Press, Inc., 10 McGovern Avenue, Lancaster, Pennsylvania.

School Review. Department of Education of The University of Chicago. Chicago, Illinois.

Composed of a series of articles which center on a single topic or general area. While the material presented is not a report of research, it does offer some of the best writing in the field on the particular issues discussed. Besides the articles, there are sections devoted to book reviews, new books in the field, and editorials.

Published quarterly: spring, summer, autumn, and winter. Single issues are obtainable by writing to The University of Chicago Press, 5750 Ellis Avenue, Chicago, Illinois.

MONOGRAPHS

Materials produced in pamphlet form are convenient to read, file, and use. Researchers can select the particular titles which reflect their interests and ignore titles which seem not to deal with what is wanted. This quality represents one large advantage of monographs which books and many journals do not share.

What Research Says to the Teacher. (A Series of Monographs). Department of Classroom Teachers and American Educational Research Association, NEA. Washington, D.C.

The purpose of the pamphlets is to keep the classroom teacher abreast of current research in the field of education. More than thirty-three separate publications are available on a wide variety of topics. Opinions on various studies are included in the reviews. At the end of each pamphlet there is a list of research references for further study.

Single copies of the *What Research Says* series may be obtained by addressing communications to the National Education Association, 1201 Sixteenth Street, N.W., Washington, D.C.

Research Resume. California Teachers Association. Burlingame, California.

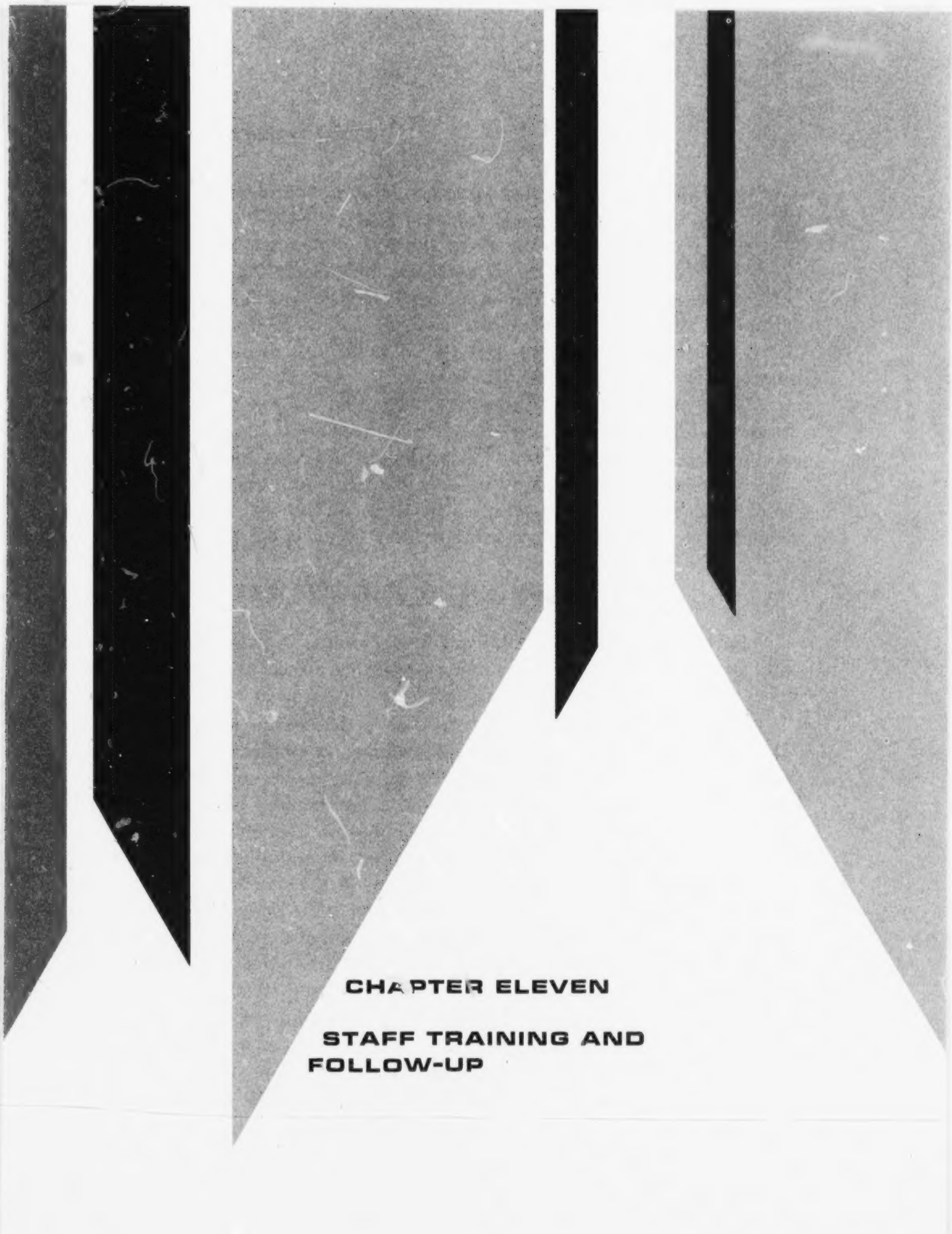
These research bulletins are intended for use in California and apply mainly to situations in that state. However, some of the topics treated are of a more general nature and hold significance for all. The subjects are of a fairly concrete nature and make for easy reading and understanding.

The publication has no set dates for issuance, but single copies will be available by title through the Publications Supply Department, CTA, 1705 Murchison Drive, Burlingame, California.

One troublesome problem that an investigator faces is to preserve in convenient form the information he has acquired from his reading. Of course, his reading will be highly selective and account for only a small part of the voluminous amount available. But even this selective part will be cumbrous to retain in memory.

The solution is to devise a suitable system. One flexible approach to the keeping of records is that of recording titles, authors, references, sources, and notes on index cards, which are then filed according to some system. The cards may be filed alphabetically by name of authors. This does not help to locate cards dealing with specific subjects, so it may be preferable to file cards according to subjects, alphabetically arranged. Or each item may be catalogued alphabetically under the name of the school system or university where the work was done. In this system, the card references from a single research group are kept together, regardless of the author. Arrangement by broad subject categories, between which there are few overlaps, is usually desirable, whatever system is employed within each category.

All in all, the major point is that the investigator should keep his records and notes in some fashion which will be easy and convenient for his work. The individual has to decide how extensive a system he wishes to use. Once decisions have been made, even a small amount of effort spent in keeping notes will seem justified.



CHAPTER ELEVEN

STAFF TRAINING AND FOLLOW-UP

It would be a mistake to assume that the reading of this book—or any other book, for that matter—would equip educational practitioners with all of the ideas and skills necessary to productive research. A book can be stimulating and informative, packed with studied and tested ideas. But it remains, after all, an inert thing without life, blood, and judgment. There is quite a chasm between ideas put in print and the effective use of those ideas in the complex and demanding world of reality around us.

The translation of a theoretical, scientific approach into warm and human educational practices may easily represent the most rewarding and inventive challenge of all. The administrator and teacher who would take the next step of finding ways to convey research ideas and techniques from a book to everyday practice need an additional set of suggestions which have been tested in actual situations with on-the-job school people.

SOME EXPERIMENTS IN RESEARCH TRAINING

Fortunately, there have been a few attempts to experiment with this problem, and the results have been promising. Two of these trials have been selected as background referents for this chapter.

The New Jersey conferences. The first experiment was comprised of a series of ten two-day research training conferences conducted between April 1953 and May 1954 for sixteen public school people from six New Jersey public school systems. The conferences were sponsored by the Horace Mann-Lincoln Institute of School Experimentation at Teachers College, Columbia University, and resulted in a publication which presented detailed descriptions and evaluations of the procedures employed.[16] These conferences were primarily concerned with human relations problems such as interactions between consultants and participants and between administrators and teachers. Relatively little stress was placed upon technical learnings such as research design and statistical analysis of data.

The Illinois seminars. The second experiment consisted of a series of three three-day research training seminars between November 1958 and February 1959. More than 250 superintendents, principals, supervisors, teachers, and school board members from public school systems in Illinois participated in the three seminars. Consultant teams were composed of college and public school research specialists. The seminars were sponsored by the Illinois Curriculum Program, a branch of the Office of the Illinois Superintendent of Public Instruction. The Curriculum Program was then housed at the University of Illinois. This experiment resulted in a doctoral dissertation report which presented extensive descriptions and evaluations of plans, results, and instruments employed.[17] Also included was a long-time evaluation of what had happened to the participants, as far as research was concerned, one year later. The seminars gave emphasis to problems in the technology of school research—problem identification, hypothesis making, research design, statistical analysis, and the reporting of results—as well as to difficulties related to interactions between consultants and learners.

Because the motives and environment for conferences and seminars away from home seemed different from what might be expected in more familiar settings, the Illinois Curriculum Program staff members conducted volunteer follow-up research seminars for approximately forty administrators and teachers in two Illinois school systems. The seminar sessions were held once each month over one school year. Results were comparable to those of the original seminars, and in conclusion a booklet was published which contained the reports of several statistical studies conducted in local schools of the two systems.[18]

PLANNING FOR LOCAL RESEARCH TRAINING

Although much remains to be learned concerning the in-school research training of administrators and teachers, sufficient guidelines, techniques, and innovations have been demonstrated and evaluated to offer valid patterns for individual school systems and individual schools to adopt and modify for their own use. This chapter focuses on the need for help from outside sources in staff training; points to where such help might be secured; describes the nature of efficient training programs; emphasizes the intellectual climate of research inquiry; and makes a proposal for long-range studies in American schools.

Help from "outside." The individual school probably comes closest to being the laboratory for learning, and it is here that theories can be tested against actuality and research results disseminated to others. The individual school can exercise more direct control over problems, subjects, and materials than can more distant institutions such as state, county, or even system-wide organizations. And the individual school is many times closer to the learning laboratory idea than are the national projects which are currently conducting curriculum studies and making recommendations.

But in spite of the individual school's advantaged position, sustained help is needed if staff members are to develop learnings and skills related to the technology of experimentation. Probably few elementary schools (and high schools, too, for that matter) are staffed with a principal and teachers who are already equipped to undertake inquiry, experimentation, and research. The necessary attitudes, knowledge, and skills can be developed as a regular part of the job through a deliberate program of research training seminars which take priority over other in-service tasks for a school year or longer. Training sessions should be held as frequently as feasible, and certainly more often than once a month.

Of course, nobody would expect an individual school staff to start from scratch in assembling and organizing all of the materials and processes desired for a sustained series of seminars. But an outside expert or experts might be obtained to work with the staff and with the principal at the same time. The whole process of committing a staff to research learnings becomes a kind of deliberately introduced and wholesome Hawthorne effect, followed by hearty reinforcement, at least during the period of training.

The principal should participate actively and directly in research processes. To become an expert leader in the research function, he will perfect his skill when he has undergone some systematic instruction with his teachers, has engaged in research study, and has learned to apply his knowledge in research operations.

Sources of help. The first source of help may consist of the periodical literature which reports complete educational research projects. The preceding chapter was given over to the identification of these sources. A school might profitably subscribe to magazines such as *The Journal of Educational Research* and the *American Educational Research Journal* for examples of current research topics and procedures.

The second source includes the locating and obtaining of consultants to help in staff training and subsequent research activities. Consultants may be found in colleges and universities, in adjacent school systems, in state offices of education, and sometimes among the psychological and guidance personnel of one's own school system. Larger school systems such as New York, Chicago, Philadelphia, Pittsburgh, and Los Angeles have long employed full-time directors of research and staff members attached to the director's office. Within the past few years, new positions designated as "director of research" have been appearing in many smaller school systems located in cities of about 100,000 population. The central office in a large enough school system can assist in the provision of consultant help and training and sometimes can cooperate with adjacent school systems and schools.

Regardless of the source for a consultant, it is crucial to obtain the *right* individual for jobs of this sort. There have been instances where a statistical researcher, competent in his own right, inadvertently quenched more fire than he fanned among naturally insecure beginners. This can be disastrous. The Illinois Curriculum Program looked for and found consultants who met three basic criteria. The criteria called for consultants who were:

1. Experienced educators, knowledgeable with respect to both theory and practice in schools
2. Reputable producers of research that might be expected to contribute to improved educational practice
3. Skillful and sensitive persons in the areas of human relations and group climate.

Individual school groups owe such careful pre-thought both to themselves and to the potential consultants. The consultant should be acquainted with what is expected from him and what is needed by the particular principal and school staff when he is invited. Criteria to aid in finding the kind of person needed should be worked out well in advance if the school's objectives are to be served.

We cannot leave the matter of subscribing to research journals and securing consultants without considering the necessity of having access to funds budgeted for research and development. Expenditures for these purposes represent legitimate uses of monies formerly taken for granted in school activities such as institutes, workshops, lectures, and the like. As school systems become accustomed, as has American industry, to improving their product through research, regular expenditures for this purpose will be regarded as quite ordinary.

The nature of training programs. In the past, school administrators and supervisors have tended to use their own graduate courses as the model for staff in-service programs. Thus, lectures and speeches by specialists, readings followed by discussions, and demonstration teaching programs have composed the usual range of techniques employed to bring information to school staffs. In other words, the techniques for in-service education activities have closely followed the traditional didactic college pattern of course offerings and direct teaching of "course" content.

But direct teaching of this sort has been recognized as cumbrous and inefficient for some types of learnings. The first obstacle introduced by college style instruction has to do with the amount of time required for learnings to mature. The second obstacle encountered in this essentially verbalistic approach to learning concerns the failure to translate course content into immediate behavior and action.

During World War II, the military forces developed a workable and efficient kind of training device which greatly speeds the time required for new learnings to take hold and which starts instruction close to the level of desired behaviors. This device is known as *simulation*. It has been employed successfully in training programs by business and industry and in certain experimental college curricula where behavioral types of learning are desired.

The Illinois Curriculum Program developed and tested a kind of simulated training that was used as the vehicle for the research training seminars which were compressed into three-day time spans. A training device from the armed forces offered leads for the seminars. During the war, the services operated a so-called dry-run for training military men in battle conditions, where the dangers of getting shot and the hazards of real shooting were eliminated. Adapting this idea, a "dry-run" training device in research practices was developed which provided a setting for practice and skill development in a threat-free and psychologically safe environment where mistakes are nontoxic. This device stressed *whole person* learning where participants could feel, act, and think the learning experience rather than just listen or talk about what they were learning. Small groups of teachers and administrators, with consultant teams, were organized into research groups to perform "thimble-size" research projects on the spot. The participants used each other as "samples" and employed synthetic, but real enough, research problems which interested them. They identified a problem, stated hypotheses, chose a research design which fitted their problem, selected a statistical test, collected the data, statistically analyzed them, came to decisions, and received immediate feed-back in terms of group critiques.

All of this was done in a "not-playing-for-keeps" atmosphere where freedom to admit one's own lack, and then learn, were prized. In the beginning stages of new, unfamiliar, and sometimes threatening learning episodes, a "dry-run" experience not only permits but also encourages uninhibited learning and skill development. Anyone participating in a simulated learning situation knows that he is only practicing and that real people and real events will be neither harmed by his ignorance nor helped through his genius. Later, after confidence and courage have been built through demonstrated successes and generous reinforcement from the trainer (consultant), a rather rapid transition can be made to real research problems involving real people in real learning situations. This brief description of training through simulation contains many implications for the sort of consultant needed by principals and teachers.

Nurturing the intellectual climate of research inquiry. Repeatedly throughout this book, we have adopted the bias that research represents an intensely personal commitment; it must belong to the researcher. Thus, the independent questions and doubts of practitioners, who are the presumed "consumers" of practices recommended for school use, form one quite necessary origin for studies designed to check and demonstrate the worth *at home* of popularized and generalized innovations. It might be said that practices hitched to recently esteemed approaches to teaching—such as phonics, modern mathematics, programs for the gifted, ability grouping, televised instruction, enrichment, science fairs and competition, language laboratories, teaching machines, departmentalization for elementary schools, et cetera, et cetera—should not be precipitously adopted *in the absence of many original research studies, and replications of research studies, performed in the individual schools contemplating revised teaching.*

Of course, popular and relatively recent innovations like those mentioned in the preceding paragraph represent only one source for needed studies. They are listed because of their currency and implied insistence and because they confront schools with pressing and immediate problems. They are *not* listed to imply a resistance to innovations in general or to suggest a retreat to the status quo. Innovations call for different kinds of thinking and are irritating enough to be generally beneficial. But if their effects are sometimes good, they may become powerful when selectively tested and followed through into practice on the basis of local research trials. On the other hand, the habit of conducting local pre-tests and research trials before committing a school to innovations, qua innovations, undoubtedly would result in throwing out a proportion of the "new" practices on the educational horizon.

The usual and customary practices which have become part of the baggage carried along by any school staff form another necessary point of origin for studies designed to check what may be too comfortably accepted and to help invent fresh "bright ideas" which have not occurred to others. In their own private worlds, many principals and teachers may have long "known" better ways to teach almost anything to almost anyone. Their secret knowledge may have been kept buried (and probably was) because they lacked acceptable ways to develop and verify it. With the gaining of knowledge and skill in research and development, this private world of truly experimental research can be widely shared with colleagues and others interested in public schools.

It is tempting in a book such as this to organize and list groups of potential studies deemed to be beneficial for someone to pursue somewhere.

But this temptation soon deflates itself because good studies begin with good problems and good questions which have been identified by individual researchers who care enough to chase them through the rigors of careful research. Practitioners can be helped to learn and acquire insights and skills in relation to the methods and techniques for conducting research studies, but selecting what is *worth investigating* from an endless number of possible choices and deciding upon creative ways to investigate it, is a quite personal matter that really cannot be done by an outsider. In effect, this process of selecting and deciding provides the rather precious intellectual climate within which research inquiry flourishes.

A proposal for long-range research studies. While the research process itself is essentially creative, personal, and fragile, the continuing educational job to be done through research obviously is heavy, huge, and almost overpowering. This imbalance between the needs and the functions of people in research becomes a major problem on the national scene. To illustrate the large dimensions of improving educational practice in American schools, we can look at some selected 1963-64 statistics which indicate the magnitude of the enterprise.

- *In 1963-64 in America there were about 82,000 public elementary schools and 25,000 public secondary schools.*
- *Enrolled in these schools were 41.8 million pupils (increasing at the rate of 1 million pupils per year for the last 11 years).*
- *These 41.8 million pupils were taught by 900,000 teachers in the public elementary schools and 650,000 teachers in the public secondary schools, for a total of 1,550,000 teachers.*
- *Expenditures for this enterprise totaled $17 billion, making public education one of America's greatest businesses.*[19]

In the past to manage and improve this "big business," we have improvised with minimal expenditures for research and development of any sort and only a tiny handful of people who professed to be researchers. And these people have been located mostly in colleges and universities—not in the public schools where learning takes place.

This state of affairs has put a stamp on educational research and has given it certain general characteristics. Educational research studies tend to be fragmentary and discontinuous. They tend, for the want of research personnel, to be performed once and never again. Their empirical tests or ideas tend to be located in one school or school system, in one part of the nation, to be assumed as valid for schools or classrooms everywhere.

The task is sheerly great in size and daring. To meet the challenge, we had best think in terms of *massive* on-going research studies in numerous situations across the country, supported by indigenous local funds. When we think of pressing educational problems being attacked simultaneously in a thousand school districts located in all parts of the country, rather than in one school system adjacent to one university, we can get a feeling for what research potentially can do.

To accomplish anything like the shift which is suggested here, two changes in our thinking on research seem obvious. The first concerns the need for greatly increasing the number of people who engage in research. Certainly, the pool of more than one and one-half million administrators and teachers in American schools is potentially rich and inviting. These are the people immersed in the natural laboratories for educational experimentation. If we have the wit to match them with the college and university research specialists, who have the knowledge but lack the numbers and the opportunities, the manpower problem can be resolved. The second change concerns the financing of research and development in local school systems and schools. When regular school funds can be annually budgeted for research, then expectations can be seriously evolved. Research grants from the government and private foundations, for the most part trickled through universities, seem to exert a limiting influence on the research that does get done. Unfortunately, the grantors generally insist that certain kinds of research projects, focused on certain research areas, are the ones worthy of support. Studies must be highly generalizable, and local needs are devalued. Implicity, this sort of bias rules out numerous projects that local people would be eager to pursue.

When local school people can gain recognition plus additional financial remuneration for their contributions to the science of education, then a motive force may be put in motion which will carry itself. When school practitioners in large enough numbers can join as equals the researchers in education and conduct their close-to-home type of experimentation, and when financial support for research permits practitioners to do the sorts of studies that they really *want* to do in many schools, then we may be in that favorable position to develop a massive attack on our best problems.

FOOTNOTES

1. National Education Association, Project on Instruction. **The Principals Look at the Schools: A Status Study of Selected Instructional Practices.** Washington, D. C.: the Association, 1962. 76 pp.

2. Royce, Josiah. **The Spirit of Modern Philosophy.** Boston: Houghton Mifflin Co., 1892. p. 72.

3. Huxley, T. H. "We Are All Scientists." **A Treasury of Science.** (Edited by Harlow Shapley, Samuel Rapport, and Helen Wright.) New York: Harper & Brothers, 1946. p. 15.

4. Tyler, Ralph W. "The Contribution of the Behavorial Sciences to Educational Research." **First Annual Phi Delta Kappa Symposium on Educational Research.** Bloomington, Indiana: Phi Delta Kappa, 1960. pp. 56-57.

5. Adapted from the class notes of David R. Krathwohl, Director, Bureau of Educational Research, Michigan State University, East Lansing, Michigan. (By permission.)

6. Kirk, Samuel A. **Early Education of the Mentally Retarded.** Urbana: University of Illinois Press, 1958. 216 pp.

7. **Ibid.**, pp. 10-11.

8. Krumboltz, John D., and Bonawitz, Barbara. "The Effect of Receiving the Confirming Response in Context in Programmed Material." **The Journal of Educational Research** 55: 472-75; June-July 1962.

9. Roethlisberger, F. J., and Dickson, W. J. **Management and the Worker.** Cambridge: Harvard University Press, 1939.

10. Corey, Stephen M. **Action Research to Improve School Practices.** New York: Bureau of Publications, Teachers College, Columbia University, 1953. pp. 134-39.

11. Association for Supervision and Curriculum Development. **A Look at Continuity in the School Program.** 1958 Yearbook. Washington, D. C.: the Association, a department of the National Education Association, 1958. 307 pp.

12. **Ibid.**, p. 287.

13. Auld, Jo Taylor. "The Grouped and the Ungrouped." **Education Report** 4: 1-2. Columbia: University of South Carolina, February 1961.

14. Moses, Lincoln E. "Non-parametric Statistics for Psychological Research." **Psychological Bulletin** 49: 122-43; 1952.

The advantages and disadvantages of non-parametric methods, as cited by Moses, are quoted in **Handbook of Social Psychology**, Vol. 1. (Edited by Gardner Lindzey.) Cambridge, Massachusetts: Addison-Wesley Publishing Co., 1954. p. 312.

15. National Education Association, Project on Instruction. **Schools for the Sixties.** New York: McGraw-Hill Book Co., 1963. p. 119.

16. Passow, A. Harry; Miles, Matthew B.; and Corey, Stephen M. **Training Curriculum Leaders for Cooperative Research.** New York: Bureau of Publications, Teachers College, Columbia University, 1955. 158 pp.

17. Patton, Earl D. "Training Public School Instructional Personnel in Research Methods." Unpublished doctoral dissertation, University of Illinois, 1962. 287 pp.

18. Barnes, Fred P., and associates. **Practical Research Projects.** Springfield: Illinois Association for Supervision and Curriculum Development, 1960. 56 pp.

19. National Education Association, Research Division. **Research Bulletin No. 42.** Washington, D. C.: the Association, February 1964. pp. 3-7.

INDEX

Amidon school, 4, 40
Auld, Jo Taylor, 69, 138

Barnes, Fred P., viii, 138
Binomial test, 87-88
Bonawitz, Barbara, 59, 138

California Achievement Test, 70
Chi square test for more than two samples, 104-105
Chi square test for two independent samples, 95-98
Chi square test of significance of differences, 35
Clelland, Richard, 77
Comparison: experimental and control groups, 53
Conant, James B., 6
Corey, Stephen M., 64, 138

Decision theory: determining levels of significance, 80-81
Dewey, John, 6, 30
Dickson, W. J., 61, 138
Dixon, Wilfred J., 77
Downie, Norville M., 77
Dual Progress Plan, 4

Einstein, Albert, 14
Experimental design: three types of evidence, 44
Experimental designs, types of, 53-73
  "after-only" with control group, 58-59
    advantages and limitations, 60-61
  "after-only" without controls, 65
    advantages and limitations, 67
  "before-after" with a single group, 61
    advantages and limitations, 62-64
  "before-after" with one control group, 53-54
    advantages and limitations, 55-56
  ex post facto experiment, 68-69
    advantages and limitations, 71-72
  extensions of control group design, 56-58
Experimental research: characteristics, 13-18

Ferguson, George A., 77
Frequency distribution control, 46
Friedman Two-Way Analysis of Variance, 98-100
Fundamental statistical problem, 75-76

Grouping: ability, 4, 70-71, 101-103

Hawthorne effect, 62
Heath, Robert W., 77
How We Think, 30
Huxley, Thomas Henry, 10, 138
Hypothesis: acceptance or rejection, 32
  null, 15, 31-32
    area of rejection, 83
    discrediting the, 79-80
    logic of, 49
  pivotal point, 30-31
  research, 31
  role of, 30

Illinois research training seminars, 130-131
Instructional television, 3

Kirk Samuel A., 54, 55, 58, 63, 138

Krathwohl, David R., 138
Krumboltz, John D., 59, 138
Kruskal-Wallis One-Way Analysis of Variance, 101-103

Lindzey, Gardner, 138
Long-range research studies, 136-137

Mann-Whitney U test, 92-93
Massey, Frank J., Jr., 77
Mayo, Elton, 61
Measurement scales: nominal, ordinal, interval, ratio, 78-79
Median test, 93-95
  extension of the, 103-104
Metropolitan Readiness Test, 31
Miles, Matthew B., 138
Moses, Lincoln E., 76, 138

New Jersey research training conferences, 130
Nonparametric methods: advantages, 78
  disadvantages, 76-77
Null hypothesis: area of rejection, 83
  discrediting the, 79-80
  logic of, 49

One-sample Chi square test, 88-89
One-tailed tests, 83

Passow, A. Harry, 138
Patton, Earl D., 138
Phi Delta Kappa: Symposium on Educational Research, 16
Population: defined, 34
  determining the nature of, 36-37
  infinite and finite, 34-35
Probability: meaning of, 48
  values, 50
Problems: not researchable, 27-28
  researchable, 28-29

Reasoning, ways of, 18-22
  concurrent use of logical-verbal and statistical, 21-22
  logical-verbal, 18-19
  statistical, 20-21
Research inquiry: intellectual climate for, 135-136
Research literature: abstracting journals, 118
  books, 114-116
  encyclopedias, 116
  general journals, 124-127
  guides, 116-117
  monographs, 127-128
  research journals, 119-124
  review journals, 119
Research process: seven steps in, 24-25
Research report: outline for, 109-110
Research types: experimental, 13
  historical and surveys, 13
Review of Educational Research, 3
Roethlisberger, F. J., 61, 138
Royce, Josiah, 9, 138

Sample: defined, 34
  in schools, 39-40
  random, 38
  stratified, 39
Siegel, Sidney, 77, 87
Sign test, 50-51, 89-92
Significance: determining levels of, 80
Simulation: in training of staff, 134
Sources of help in staff training, 132-133
Spearman Rank Correlation Coefficient, 105-107

Statistical methods: principal problems, 48
Statistical models: parametric and nonparametric, 37
Statistical significance: levels, 50
  meaning of, 46, 49
Statistical tests: nonparametric, 76
  parametric, 76
Statistics: descriptive, 20, 34
  inferential, 20-21, 34
  probability, 21

Tate, Merle W., 77
Teaching machines, 3
Two-tailed tests, 83
Tyler, Ralph, 16, 138

Variable: dependent, or criterion, 43
  independent, or experimental, 43

What authorities say, 6-7
What everybody knows, 5-6
What practitioners doubt, 9-10
What practitioners feel, 8-9
What practitioners think, 10-11
What researchers say, 7-8